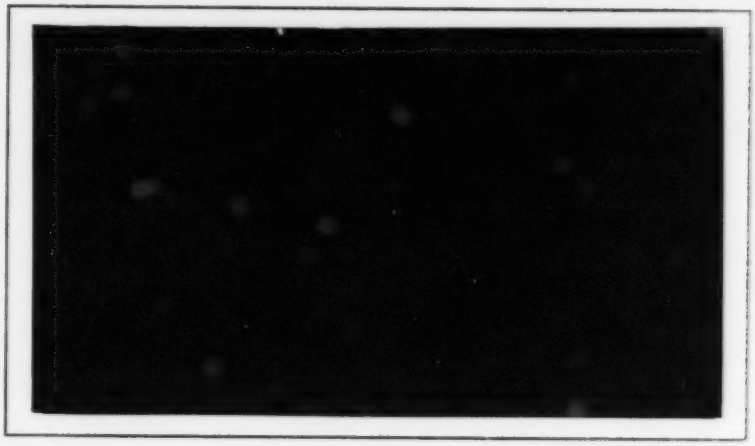

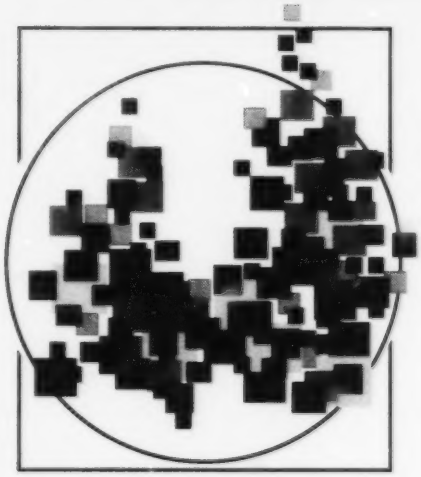

CENTRE FOR
SYSTEMS
SCIENCE

Simon Fraser University, Burnaby, B.C., Canada V5A 1S6 / (604) 291-4704

92-05742/

A Sense-Based, Process Model of Knowledge and Belief

Robert F Hadley

CSS-IS TR 91-02

Substantially revises and modifies CSS-IS TR89-04
January 1991

A Sense-Based, Process Model of Knowledge and Belief

by

Robert F. Hadley

School of Computing Science

Simon Fraser University

Burnaby, Canada V5A 1S6

Tel. 604-291-4488

hadley@cs.sfu.ca

Copyright © 1991 Robert F. Hadley

ABSTRACT

A process-oriented model of belief and knowledge is presented which permits the representation of nested propositional attitudes within first-order logic. The model (NIM, for 'nested intensional model') is axiomatized, sense-based (via intensions), and sanctions inferences involving nested epistemic attitudes, with different agents and different times. Because NIM is grounded upon senses, it provides a framework in which agents may reason about the beliefs of another agent while remaining neutral with respect to the syntactic forms used to express the latter agent's beliefs. In addition, the axioms of NIM may be used to *justify* conclusions which an agent derives concerning the beliefs of another agent. However, because NIM makes explicit many of the process-related preconditions of belief acquisition, NIM underscores the limitations of inter-agent reasoning. Moreover, NIM provides agents with a conceptual map, interrelating the concepts of 'knowledge', 'belief', 'truth', and a number of cognate concepts, such as 'infers', 'retracts', and 'questions'. The broad scope of NIM arises in part from the fact that its axioms are represented in a novel extension of first-order logic, $\mathcal{T}$-FOL (presented herein). $\mathcal{T}$-FOL is a logic which simultaneously permits the representation of truth ascriptions, implicit self-reference, and arbitrarily embedded sentences within a first-order setting. Through the combined use of principles derived from Frege, Montague, and Kripke, together with context-sensitive semantic conventions, $\mathcal{T}$-FOL captures the logic of truth inferences, while avoiding the inconsistencies exhibited by Tarski. By virtue of a distinguished truth predicate, together with inference rules involving this predicate, and its context-sensitive semantics, $\mathcal{T}$-FOL attains an expressive power approaching that of natural language. Applications of $\mathcal{T}$-FOL and NIM to inter-agent reasoning are described and the soundness and completeness of $\mathcal{T}$-FOL are established herein.

1. INTRODUCTION

Within the cognitively related disciplines (including AI) there are three dominant approaches towards the modeling of epistemic states.[1] For convenience I call these the logistic, syntactic, and intensional approaches. The *logistic* approach seeks to discover a *logic* of belief (knowledge) which would enable one to derive, *a-priori*, the totality of things an agent believes (or knows) at a particular time, given the formulae explicitly written in the agent's belief base.[2] The *syntactic* approach (as Levesque, 1984, dubs it) identifies an agent's beliefs with just those formulae explicitly written in the agent's belief base -- the agent has no other beliefs. By contrast, the *intensional* approach identifies beliefs with abstractions, e.g., with *semantic contents* or propositions. On this account the same belief may be expressed by means of a limited variety of distinct syntactic objects. As will emerge in what follows, our present concerns lie with the intensional model, and with how intensions (more exactly, *senses*) are related to the *processes* of inference, belief acquisition, and belief retraction. In section 2, I present a first-order logic, $\mathcal{T}$-FOL, which permits the arbitrary nesting of sentences within propositional attitudes, and indeed, within any predicate. $\mathcal{T}$-FOL's semantic conventions are *context-sensitive*, thereby achieving implicit self-reference and arbitrary embedding of sentences without the encumbrance of quotation marks and concatenation devices. In addition, $\mathcal{T}$-FOL includes a distinguished **truth predicate** and special inference rules which permit inferences involving truth ascriptions to be drawn in a natural, intuitive fashion, while avoiding contradictions of the kind noted by Tarski (1936).

In section 3, I present an axiomatic, *nested intensional model* of belief (NIM), which is described in the language of $\mathcal{T}$-FOL. Unlike most *epistemic logics* found in AI, NIM aims to provide a theory of *belief-change*, which is sensitive to the empirical contingencies of epistemic *processes*. While NIM sanctions inferences among nested epistemic propositions, it does not purport to be a logic of belief (in the sense described below).

Section 4 presents an application of NIM and $\mathcal{T}$-FOL to *inter-agent* reasoning. An example

---

[1]Portions of this paper were presented to the Symposium on Computational Intelligence (Milan, 1990) and the Tenth Annual Conference of the Cognitive Science Society (Montreal, 1988), and appear in the proceedings of those conferences.

[2]Hereafter, "belief" shall be taken to include knowledge, unless the context requires that the distinction be explicit.

illustrates the manner in which an agent may make plausible inferences about *some* beliefs another agent is likely to acquire, without requiring the former agent to know the representation language of the latter agent.

In the remainder of this Introduction, we briefly examine the Logistic and Syntactic approaches to belief, and contrast these with the Intensional approach adopted here.

1.1 The Logistic Approach

Those who adopt the logistic approach are concerned with the discovery of *inference rules*, such that an agent's explicit beliefs are *closed under* the application of these rules. Typically, the epistemic logics proposed do not require an agent's beliefs to be closed under *all* valid inference rules (full logical omniscience is not required), but closure under *certain* inference rules is still required. Thus, these logics typically require that agents each possess an infinity of beliefs, or at least believe an infinity of distinct formulae. (cf. Levesque, 1984; Halpern & Moses, 1985; Lakemeyer, 1987). Some have questioned the motivation for introducing logics which represent agents as believing an infinity of formulae (e.g., Fagin & Halpern, 1985; Hadley, 1991), but the underlying rationale may be that such logics employ a non-standard sense of belief, which yet has application in the context of finite agents. For example, the sense of belief appropriate to (Levesque, 1984) *might* be such that "*p* is believed" means approximately that one could *eventually come to believe p* by a tractable computational process.[3] However, whether or not one thinks that epistemic logics (of the "infinite belief set" variety) require us to adopt a non-standard sense of belief, such logics usually are not concerned with certain important *process-related* aspects of belief. For example, with limited exceptions[4] such logics do not address the fact that an agent acquires and *changes* beliefs over time, or that these processes often involve inference, testing beliefs for consistency, retraction, etc. Doubtless, most epistemic logics were never intended to address those aspects of belief, but for certain applications one desires a *formal model* of how these cognitive processes relate to the belief states of the agent. An agent *X* who knows what another agent *Y* believed a minute ago may wish to form a reasonable conjecture about what *Y* believes

---

[3] I am indebted to Jim Delgrande for suggesting this sense of 'belief'.

[4] Konolige, (1985) presents an epistemic logic which does address the process of belief *introspection*. However, Konolige's system is not intended to address most other process-related aspects of belief. See also, Martins & Shapiro (1988), discussed presently.

now. If so, *X* requires at least a crude process-theory of belief to arrive *even* at a reasonable conjecture about *Y*'s present beliefs.

Of course, some aspects of belief change are addressed by truth maintenance systems (e.g., de Kleer, 1986; Doyle, 1979), but TMSs are concerned with *procedures* for manipulating belief sets, rather than with formal models which interrelate the concepts of belief, knowledge, inference, retraction, etc. By contrast, Martins & Shapiro (1988) describe a *logic* for belief revision which formalizes the process of recording inferential dependencies, and isolating "culprits" in contradictory belief sets. However, Martins and Shapiro focus only upon a subset of the issues which concern us because they are concerned with the *details* of how beliefs are revised in the face of contradictions. The present model, NIM, notes that belief revision (in the face of contradiction) is one of several ways in which a belief set may change. NIM also addresses the issue of synonymous sentences expressing the same belief, and the fact that *knowledge* implies truth. Also, as mentioned, $\mathcal{T}$-FOL addresses issues of representation and inferences involving truth predicates.

1.2 The Syntactic Approach

Now, apart from epistemic *logics* which posit an infinity of beliefs, certain *syntactically-oriented* belief models address the fact that beliefs can be acquired and retracted by *processes*. Both Goodwin (1985) and Doyle (1983) present process theories for those non-monotonic reasoning systems which involve the use of a TMS. However, in both these theories the assumption is made that any agent who infers a given conclusion necessarily comes to believe that conclusion (at least temporarily). This assumption appears to preclude the possibility that a sophisticated agent may choose to investigate whether the conclusion conflicts with other long-standing beliefs, before "committing" to that conclusion. Moreover, the process theories of Goodwin and Doyle share a crucial aspect with other syntactic approaches to belief (e.g. Hass 1985, 1986; Maida, 1986), namely, they identify beliefs with particular syntactic objects, so that, for example, an agent might believe the formula $p \vee q$ but not $q \vee p$, or believe "Tadpoles swim" but not "Pollywogs swim". Now, many reject the syntactic approach entirely, on the grounds that it is too fine-grained. The syntactic approach arguably fails to account for the fact that *some* syntactically distinct formulae seem to express precisely the same thought, and therefore ought to have the same belief status (cf. Church, 1954; Levesque, 1984; Hadley, 1986). No doubt, there are particular applications where one can safely ignore the fact that syntactically distinct formulae sometimes express the same

belief (cf. Hadley, 1991), but our present goal is to provide a model of epistemic processes which is *general* in scope, encompassing both human and artificial agents.

Before leaving this discussion of purely syntactic epistemic models, we should note that Fagin and Halpern (1985) present what might be viewed as a *liberalized* syntactic belief model. The Fagin-Halpern (F-H) model identifies an agent's beliefs with a *subset* of formulae of which the agent is aware. This particular subset (which satisfies conditions imposed by a specified epistemic logic) *may* be taken to include formulae the agent will only *come to be* aware of at some future time. In this respect the F-H model is less restrictive than those which restrict an agent's beliefs to formulae presently written in the agent's belief base. However, apart from the awareness requirement, the F-H model is not concerned with the empirical preconditions of belief (e.g., that a sophisticated robot might derive a formula from existing beliefs, but not "commit to" the formula until a "consistency check" with other beliefs is complete).

There is no reason to suppose the F-H model was ever intended to deal with empirical preconditions of belief acquisition and change, but an agent who wishes to reason (or conjecture) about what a given robot *currently* believes will require a model of belief which makes explicit many of these preconditions. Such a model will need to recognize the fact that agents revise and retract beliefs, and that they sometimes derive formulae which they *never* accept, even though the derived formula is indeed a valid consequence of the agent's existing beliefs.

1.3 The Intensional Approach

The intensional approach is founded upon the premise, widespread among philosophers of language, that belief and knowledge must be understood as a relation between an agent and an abstract *intension* (proposition or sense). Within AI, a number of researchers (e.g., Hobbs & Rosenschein, 1978; Barnden, 1986; Hadley, 1988) have been influenced by the intensional semantics of Montague (1970) or Lewis (1970). Others (e.g., Brachman, 1977; Maida & Shapiro, 1982; Rapaport, 1986) have focused upon the relationship of intensions to semantic networks. In the present work, I am concerned to integrate a *particular form of intensional semantics* with the first-order representation language, $\mathcal{T}$-FOL, and the formal model of epistemic processes, NIM.

The thesis that a declarative sentence expresses an abstract sense or intension, whose content is a *function* both of the sentence's syntax and of the meaning of its parts, is a view which dates back at

least to Frege (1892). Modern philosophical descendants of Frege have considerably refined and extended Frege's view along a number of dimensions, but Frege's underlying idea that the sense (intension) of a sentence is determined by a compositional process, involving both syntax and meaning functions, persists. Of those working along Fregean lines, some (e.g., Montague, 1970; Lewis, 1970) have used 'intension of a sentence' to refer to an abstract *function* which, given a possible world as an argument, returns the truth value of the sentence in that possible world. Although this usage of intension bears some resemblance to Frege's notion of *sense*, the resemblance is not complete.

As Montague and Lewis were certainly aware, intensions construed purely as functions are too coarse-grained to serve as senses in Frege's usage. For Frege, two expressions have identical senses only when they can be substituted for one another in belief ascriptions, without affecting the truth value of that ascription. By contrast, sentences which share the same intension, on Montague's and Lewis' usages, often fail the substitutivity test.[5] Nevertheless, Fregean senses resemble intensions in this respect -- the (Fregean) sense of a declarative sentence *determines* its truth value in the actual world.

Now, our purpose here is not to provide an exegesis or defense of Frege's use of 'sense', but rather to indicate the *kind* of thing we have in mind when we refer to *senses*. We require *senses* to satisfy the following conditions:

1. Senses have a granularity appropriate for individuating beliefs (distinct sentences may express the same sense and the same belief).
2. The sense of a sentence is determined by a compositional process involving the syntax of the sentence and the semantic content of its constituent parts.
3. Senses (or a component of the sense) must be capable of (functionally) determining the truth value of a sentence in a possible world. For convenience, in the event that a *component* of the sense determines truth values in possible worlds, I shall still say that "the sense determines the truth value".

The last of our three conditions is phrased so as to permit a *component* of a sense to behave as one of Montague's intensions. For example, we might regard the sense of a sentence S as an ordered

---

[5]Note that when intensions are construed as *functions* from possible worlds to truth values, every tautology has the same intension (since every tautology is true in *every* possible world). Yet, distinct tautologies cannot reliably be interchanged in belief ascriptions without affecting the truth value of the belief ascription.

pair, whose first element is an (Montague-style) intension of S, and whose second element contains syntactic, semantic, and/or pragmatic information about S, sufficient to provide the requisite granularity. Lewis (1970) provides a definition of *sentence meaning* which might be viewed as conforming to this *ordered pair* pattern. However, some reject Lewis' account of sentence meaning as excessively fine-grained, since Lewis allows that sentences are *synonymous* only when their intensions are identical *and* their parse trees are identical. Nevertheless, as noted in (Hadley, 1988), a minor revision of Lewis' account may provide a granularity suitable for substitution in belief contexts. We merely stipulate that two sentences have the *same sense* when their intensions are identical and their parse trees are isomorphic in a graph-theoretic sense. For example, on this account $p \vee q$ and $q \vee p$ have the same sense, as do 'Tadpoles swim' and 'Pollywogs swim'. On the other hand, $p \wedge (q \vee r)$ would not have the same sense as $(p \wedge q) \vee (p \wedge r)$, since these sentences are structurally non-isomorphic.

Other *compositional* theories of sentence meaning exist, of course, some of which postulate semantic objects which may satisfy our criteria for 'sense' (cf. Barwise & Perry, 1983; Cresswell, 1985). Such semantic objects are variously named (e.g., 'structured meanings', 'propositions', 'interpretations', 'semantic contents') and important differences exist among them. Nevertheless, provided these objects may be regarded as compositional, appropriately fine-grained (for belief), and capable of determining truth (in a functional sense) I shall regard them as senses.

. . . . . .

The plan for the remainder of this paper is as follows: in section 2, I present the language and logic, $\mathcal{T}$-FOL, which at once enables the representation of nested propositions, implicit self-reference, and reasoning with truth ascriptions. Applications of $\mathcal{T}$-FOL involving truth predicates are provided. In section 3, a sense-based theory of epistemic processes is presented, using the language of $\mathcal{T}$-FOL. This theory seeks to formalize many of the interrelationships among epistemic concepts, and related notions such as retracting, questioning, or committing to a belief. To the author's knowledge, no prior epistemic model addresses the range of issues which NIM and $\mathcal{T}$-FOL collectively integrate. An application of $\mathcal{T}$-FOL and NIM to inter-agent reasoning is presented in section 4.

2. $\mathcal{T}$-FOL, THE LANGUAGE AND LOGIC

Nested epistemic assertions and *implicitly self-referential* statements are an important part of human dialogue. Likewise, truth ascriptions, and inferences involving truth ascriptions, play a significant role. Consider the following examples:

(a) Mary: "Bill said he wasn't interested in dating Sally."
John: "Well, Bill is not always truthful about his romantic life."

(b) Sue: "Nobody's infallible; not all my beliefs are true.
On the other hand, when I tell you something about New York, it's very likely to be true.

(c) Bill: "I called because I didn't know whether you knew I would be at the party".

Example (a) contains an embedded sentence (in Mary's utterance), and John's denial of truth to some of Bill's statements. John's truth-denial may very well lead Mary to alter her attitude towards Bill, and to adopt a skeptical attitude towards several of Bill's other statements. Example (b) presents two sentences, the first of which could well be (implicitly) self-referential, since the speaker presumably counts that very sentence among her beliefs. The second sentence involves a truth ascription which could, if accepted, lead the listener to accept several of Sue's statements about New York. In (c) we are presented with an epistemic proposition (involving knowledge) nested within another epistemic assertion.

Examples such as the above form an integral part of human cognition, and "user model" components of automated systems will, of necessity, be incomplete models of human users, unless the principles underlying such dialogue can be understood in formal terms. Yet, surprisingly little attention has been devoted to the development of a logic which *at once* sanctions inferences involving truth ascriptions while *also* permitting self-reference and arbitrary nesting of epistemic sentences within other epistemic sentences. A notable prior attempt to integrate these multiple aspects of human cognition within a formal setting is Perlis' (1985) first-order treatment of truth. A minor disadvantage of Perlis' approach is his use of nested quotation marks and string-concatenation functions to achieve a first-order language. In addition, in section 2.2, I argue that Perlis' treatment of truth does *not* obviate some forms of semantic paradox which are discussed by Kripke (1975).

Now, it is well known that efforts to capture the logic of truth ascriptions in a first-order language are fraught with difficulty (cf. Tarski, 1936). In $\mathcal{T}$-FOL, these difficulties are circumvented through the integration of a Fregean-based semantics with Kripke's theory of truth. The compositional

aspects of $\mathcal{T}$-FOL's semantics are clearly inspired by Montague's (1970) work on intensional semantics. The primary contribution of $\mathcal{T}$-FOL lies in its integration of these various approaches within a single *first-order* logic. Within $\mathcal{T}$-FOL, truth inferences are sanctioned by the addition of two rules of inference to standard FOL. These rules involve a distinguished truth-predicate, 'T' (hence, $\mathcal{T}$-FOL). The soundness and completeness of $\mathcal{T}$-FOL will be established here.

2.1 The Language of $\mathcal{T}$-FOL

The language of $\mathcal{T}$-FOL extends the expressive power of classical predicate logic. Readers acquainted with the prevailing affinity (within AI) for epistemic *modal* logics may wonder why I have chosen a predicate-based language rather than a modal one. The principal reason is this: modal logics permit us to embed sentences only in the presence of certain distinguished *modal* operators. Typically, these are '*necessity*', '*possibility*', '*K*' (knows), and '*B*' (believes). Although other modal operators may be included, the introduction of each new modal operator alters the entire logic, and necessitates new soundness and completeness proofs. By contrast, within $\mathcal{T}$-FOL, embedded sentences may be written as arguments to *any* predicate we choose, without extending the logic. For example, we may formulate premises such as, *doubtful(loves(john, mary))*, *humorous*$(\exists x \exists y\,(\,teases(x, y))$, and $\forall x\,\{probable(x) \rightarrow \neg\, unlikely(x)\}$, where we may universally instantiate on any sentence or constant of the language. Moreover, as Perlis (1988) demonstrates, once modal logics are enriched to the point where general *substitution* operations are expressible, they become vulnerable to the same paradoxes as predicate logics. Finally, as Perlis also remarks, the plethora of modal logics in AI threatens to create a tower of Babel. If we insist upon expressing our theories of belief in modal formalisms, then each modification to our theory necessitates a new logic, with its own semantics. By contrast, any number of epistemic theories may be expressed in $\mathcal{T}$-FOL without affecting the semantic *conventions*. We are assured by the soundness and completeness of the logic that all and only semantically correct consequences of the theory (and universally valid theorems) are derivable. It is worth noting that *completeness* would be lost if we were to resort to a higher-order logic to achieve the embedding of sentences (assuming equal expressive power in other respects).

Now, as previously noted, Perlis (1985) describes a method, involving quotation, unquotation, and concatenation of expressions, which permits the representation of nested propositions in an extension of FOL. Haas (1986) presents a quotation-based scheme which is, in many respects,

analogous to that of Perlis. However, both the methods of Perlis and Haas become unwieldy at levels of nesting greater than one. By contrast, the language of $\mathcal{T}$-FOL involves no quotation mechanism. (Hereafter, the language of $\mathcal{T}$-FOL is called $\mathcal{L}$). $\mathcal{T}$-FOL relies instead on the *composition* of *functions* to ensure both that the objects of belief are senses, and that inferences are not unduly encumbered by a plethora of quotation marks. For example, consider the following sentence which involves self-reference within an embedded sentence.

(1) Mary knows that John believes that none of John's beliefs are absurd.

In $\mathcal{T}$-FOL (1) would be represented as

(2) $Knows(mary, Believes(john, \neg\exists\, x\{\, Believes(john, x) \,\&\, Absurd(x)\,\}\,))$

Now, at first glance (2) appears not to be a sentence of first-order logic (FOL). However, as we shall see, if we allow the role of predicates, quantifiers, and truth connectives to be *context sensitive*, so that they express functions in any embedded context, then (2) may be viewed as having the following *first-order* form:

$$Knows(mary, f1(john, f2(\, f3(\, f1(john, x))) \,\&\, f4(x)))$$

The underlying motivation for this approach is not merely to simplify our system of representation and inference (though we do obtain those benefits). Rather, we are attempting to embody Frege's thesis that the denotation of a sentence (or wff) varies according to its surrounding context. Frege maintained that sentences which occur in embedded epistemic contexts (as in (1)) denote their *sense*, and not their truth value. Our construal of this thesis shall be that *all wffs*, embedded or not, *return* their own senses through a compositional process involving the senses of their constituents. The *context* surrounding a wff determines whether the sense so returned is evaluated for truth. A non-embedded, closed wff (i.e., a sentence) has its sense evaluated, as we would expect. By contrast, the sense of an embedded wff, open or closed, is never *evaluated*, but is passed upwards as part of the compositional process.

In order to implement the context-sensitive assumptions just outlined, I adopt the following notational conventions: the predicates *T* (for true), *K* (for knowledge), and *B* (for belief) are distinguished predicates, and always appear in upper case; other predicates, constants, and variables are written in lower case. Officially, the truth-functional connectives, '$\neg$', '$\rightarrow$', etc., are *prefix* connectives whose arguments are written within parentheses. Thus, we write

$\rightarrow (\forall x\,(pred1(x)),\ \exists y\,(pred2(y)))$. Also, any time a quantifier is introduced, the top-level formula within its scope is enclosed in parentheses, as in $\exists x\,(pred2(x))$. In practice, I omit parentheses where the scope is unambiguous, and use connectives in their customary infix position. Our official notation is intended merely to stress that predicates, connectives, and quantifiers may all be viewed as functors. The functions expressed by these functors expect to be *passed* the *senses* of terms or wffs, and will return (as value) a composite sense, which includes the sense of the quantifier, connective, or predicate normally associated with that functor.

Composite senses may be formed in a variety of ways, depending upon the particular sense-theory adopted. However, in order to justify the application of quantificational rules to embedded wffs, it will be necessary to work with some *particular*, compositional theory of senses. Therefore, in the sequel, I adopt a particular approach. Bear in mind, however, that alternative approaches to senses are compatible with the overall compositional conventions we require.

To begin, I shall assume that the sense of a *standard* constant, c, is an ordered pair, $<f_c$, *sem-info*>, where f_c is a Montague style *intension*, which returns the individual assigned to c, given a possible world. *Sem-info* may be any other semantic information which we wish to associate with c (such as defeasible procedures for identifying the denotation of c, cf. Hadley, 1989).

In addition to *standard* constants (whose sense and denotation are assigned when the language is created), certain constants acquire their sense when they are introduced via existential instantiation. For example, given a premise (Σ) having the form $\exists\,\alpha(\theta(\alpha))$, we may obtain $\theta(c1)$ by instantiation. The validity of existential instantiation requires that, prior to the act of instantiation, no constraints have been placed upon the sense (and denotation) of $c1$. However, when we instantiate the sense of $c1$ is immediately constrained as follows: $c1$ acquires a sense of the form $< f_{c1}$, *sem-info*>, where f_{c1} is a function which (given a possible world) *returns* an individual that satisfies $\theta(\alpha)$, if such an individual exists in the given possible world. Otherwise, f_{c1} still returns an individual from that world, but since the individual does not satisfy $\theta(\alpha)$, $\theta(c1)$ would be false in that world (which is what we expect, since (Σ) should be false in that world.) In other words, the function f_{c1} serves as an intension for $c1$, and behaves so that $\theta(c1)$ is true just when (Σ) is true. Note that, if (Σ) is true in a given world, more than one individual may satisfy $\theta(\alpha)$. In this case, any of several functions could serve as f_{c1}. When we perform existential instantiation, it is not our task to select

among these possible functions.[6] Rather, we simply assume that *c1* acquires such a function as its intension, and comes to have the sense $<f_{c1}$, *sem-info*>. In the present case, *sem-info* may be comprised of defeasible procedures for selecting individuals which satisfy $\theta(\alpha)$. (If $\theta(\alpha)$ is unsatisfiable because θ is an unsatisfiable predicate, then *sem-info* may be the empty set.)

Having described the senses of individual constants, we now consider the *senses* of individual variables. Although it may seem odd to assign senses to variables, it will be convenient to assume that open wffs, such as '*red(x)*', have senses, and that the identity of the variable contributes to this sense. As will emerge in the sequel, this can be achieved by assuming that the *sense* of each individual variable *is itself*.

The sense of any (closed or open) wff of the form ϕ*(arg-1, ..., arg-n)*, where ϕ is a predicate, is a *tuple* of the form

(3) $< f_\phi,\ sense_{arg\text{-}1}, \ldots\ldots, sense_{arg\text{-}n},\ sem\text{-}info_\phi >$.

In (3), f_ϕ is the *intension* associated with ϕ. This (Montague-style) intension determines the truth value of all *non-embedded* sentences having the form $\phi(arg_1, \ldots, arg_n)$. (Thus, f_ϕ is determined when the extension of predicate ϕ is assigned.) Each $sense_{arg\text{-}i}$ in (3) is the sense of the *i-th* argument to ϕ, and $sem\text{-}info_\phi$ is any other semantic information we wish to associate with ϕ (by analogy with *sem-info* for constants; *nuance* information may also be included.) Instances of (3) containing no variables shall be called *closed* senses. (Recall that if arg_i is a variable, then $sense_{arg\text{-}i}$ is that same variable.) In general, any sense containing *no* variables (*at any level* of nesting) is a *closed* sense. (Closed senses may also arise through quantifier binding, as we shall see.) Closed instances of (3) may be *evaluated* for truth by first evaluating the *intensions* of each $sense_{arg\text{-}i}$ to obtain a series of objects in the domain of $\mathcal{L}$, and then applying the function f_ϕ to this n-tuple of objects. Of course, truth evaluation does not occur at embedded levels.

Senses of truth-functional (molecular) wffs have the form

(4) $< f_\psi, sense_{arg\text{-}1}, \ldots\ldots, sense_{arg\text{-}n} >$

---

[6]Analogously, when *skolem functions* are introduced to eliminate existential quantifiers when converting to *clause form*, more than one function might actually serve as the skolem function. Customarily, one simply assumes that one of these functions is selected as the skolem function.

where each $sense_{arg\text{-}i}$ is the sense of a (closed or open) wff of $\mathcal{L}$. (It sometimes happens that a given $sense_{arg\text{-}i}$ has the form of (4) itself, in which case recursive decomposition ultimately yields atomic senses.) In (4), f_{ψ} represents the truth-function associated with the truth-connective ψ. For example, the wff

(5) $\rightarrow$*(parent(john, x), child(x, john))*

has the sense

(6) $<f_{\rightarrow}, \ <f_{parent}, sense_{john}, x, sem\text{-}info_{parent}>, \ <f_{child}, x, sense_{john}, sem\text{-}info_{child}>>$.

Note that the **functor** '$\rightarrow$' expresses the same function, *f1*, whether or not the functor occurs in an embedded context. The function *f1* takes the senses it is passed, and returns a composite sense. For example, the occurrence of '$\rightarrow$' in (5) expresses a function which takes the senses of the given arguments, and return the composite sense shown in (6). The situation is analogous with quantifiers. Expressions such as '$\forall x$', '$\exists y$' are regarded as functors. The functions they express take an *open* sense as argument, and return a sense having the form $<f_{quantif}, sense_{arg\text{-}1}>$. Here, $f_{quantif}$ is the *intension* of the relevant quantifier. For example, the sense returned by the functor '$\forall x$' in

(7) $\exists y(\forall x(loves(y, x)))$

is

(8) $<f_{\forall x}, <f_{loves}, y, x, sem\text{-}info_{loves}>>$.

Note that (8) is an *open* sense, because (though 'x' may be considered bound by $f_{\forall x}$) the variable y is unbound in (8). If y were replaced by the sense of a constant, the resulting version of (8) would be a *closed* sense. So, suppose that instead of (7) we had

(7′) *incredible(* $\forall x$ *(loves(jane,x)))* .

Then the sense returned by '$\forall x$' in (7′) would be closed, and is represented thus:

(8′) $<f_{\forall x}, <f_{loves}, sense_{jane}, x, sem\text{-}info_{loves}>>$.

Since (8′) is closed, it could be evaluated for truth by *applying* the intension $f_{\forall x}$ to the remainder of (8′). Since $f_{\forall x}$ is the intension of a universal quantifier of x, $f_{\forall x}$ will return truth if and only if its argument (<f_{loves}, ..., x, ...>) evaluates to true for *all* replacements of 'x' by the sense of a constant. The intension of existential quantifiers behaves in an analogous fashion (replace 'all' by 'some').

Given the role of quantifier intensions ($f_{\forall x}$, $f_{\exists y}$, etc.), the standard quantifier inference rules (universal and existential instantiation and generalization) can be justified in a straightforward fashion. For example, (8′) has the general form:

(9) <$f_{\forall x}$, *open-sense*$_x$>.

Let us assume that *open-sense*$_x$ contains no free variables apart from x. Also, assume that a given *open-sense*$_x$ becomes grounded when all occurrences of x in that sense are replaced by the *sense* of a constant. Then, by definition of $f_{\forall x}$ we know that a given instance of (9) will evaluate to true exactly when *every* grounding of the given *open-sense*$_x$ evaluates to true. From this it follows that whenever a given instance of (9) evaluates to true, any wff expressing the given *open-sense*$_x$ would express a true sense, when occurrences of 'x' are replaced by the same arbitrary constant throughout that wff. This means that ***universal instantiation*** is justified for any *sentence* of the form $\forall x(\pi)$, where π is an open wff involving the variable x. Moreover, since the validity of this reasoning does not depend upon the identity of the *particular* variable 'x', we conclude that *universal instantiation* always yields valid results, within $\mathcal{T}$-FOL. Clearly, we could reverse the above reasoning to establish the validity of ***universal generalization***. Analogous reasoning would establish existential instantiation and generalization, respectively.

At this point it may be well to address a concern that will have occurred to some readers, namely, that each *sense* (of a wff) possesses a structure isomorphic to the syntax of a wff expressing that sense. Since the senses I have posited do indeed mirror the structure of their syntactic expressions, the question naturally arises whether this sense-based approach presents real advantages over a straightforwardly syntactic approach. The answer is that they do. To begin with, it is obvious that distinct wffs, such as '*tadpole(x)*' and '*polliwog(x)*' can express identical senses, given our concept of 'sense'. Also, in section 1.3 we considered a suggestion whereby the criteria for *identity* of senses may be relaxed somewhat, so that structurally isomorphic sentences, which are logically

equivalent (though symbolically distinct) may be *regarded* as expressing the same sense. This would enable us to say that $\exists x(P(x) \vee Q(x))$ has the same sense as $\exists y(Q(y) \vee P(y))$. Moreover, it is clear that syntactic expressions must have senses (or meanings) *of some kind* if they are to assert anything, and I have already reviewed arguments to the effect that having a belief involves a relationship between an agent and what is asserted by a sentence. Finally, as will emerge in the sequel, the assumption that truth is predicated of senses, and not sentences (except derivatively), obviates certain forms of *direct* self-reference which lead directly to paradox. For, while there seem to be only *ad hoc* reasons for barring a name from denoting a sentence which contains that name, there are principled semantic objections to allowing a name to denote a sense expressed by a sentence containing that name.

2.2 Truth in $\mathcal{T}$-FOL

Hopefully, it is now clear that the functions I have chosen to assign to predicates and logical operators are such that when a first-order sentence, ψ, is nested inside another wff, ψ *returns* (but does not evaluate) the ordinary, non-embedded sense of ψ. For this reason, we may justify the addition of two rules of inference to our first-order theory, given the convention just described. These rules allow us to move an expression into or out of an embedded context of the following form:

(10) $T(\phi)$.

Informally, we may construe T as the "truth" predicate. Thus, (10) would assert that the *sense* returned by ϕ is 'true'. A more formal interpretation for T is provided in section 2.3. At this stage, it is important to note that I do not attempt to *explicitly* define the truth predicate within $\mathcal{T}$-FOL. However, the following two rules of inference express the *logic* of this predicate, and they are included in our extension of FOL.

T1: ϕ may be derived from T(ϕ), provided ϕ has the typographic form of a (closed) sentence of $\mathcal{L}$.
T2: T(ϕ) may be derived from ϕ.

Note that T1 allows us to "extract" expressions from a truth context, under certain conditions, and T2 allows us to *embed* a sentence in a truth context. Now, there are circumstances in which we wish to assert a sentence such as '$T(c)$', where 'c' is an individual constant denoting some sense.

In such cases, T1 does *not* allow us to derive *c* as an assertion, since *c* is not a sentence of $\mathcal{L}$.

Readers will note that rules T1 and T2, taken jointly, are equivalent to the axiom schema:

$$\text{True}(\phi) \equiv \phi$$

where the above restrictions are placed on ϕ.[7] In light of Tarski's (1936) demonstration that attempts to enrich FOL -- by including a schema *somewhat* similar to the above -- rapidly leads to semantic paradox (e.g., the Liar paradox), the question naturally arises whether $\mathcal{T}$-FOL is likewise vulnerable to paradox. The short answer is that, although $\mathcal{T}$-FOL is vulnerable to certain kinds of *contingent* paradox, these do not threaten the consistency of the logic (as we shall see, presently). To begin with, it does not appear possible to adapt Tarski's argument to $\mathcal{T}$-FOL's semantics. For Tarski's argument (and Montague's (1960) variant of the argument) presupposes both that *truth is predicated* of sentences, and that sentences may legitimately contain proper names of themselves. By contrast, in $\mathcal{T}$-FOL, truth is not predicated of sentences, but of senses. Moreover, the principle of semantic *compositionality* prevents names from denoting the *sense* of a sentence containing that same *name* (because this would require the name to denote a sense which contains, as a part, the sense of that same name)[8]. The matter is more complex when *definite descriptions* are substituted for names in self-referential assertions, because a given *sense* may *contingently* satisfy a definite description. However, even when definite descriptions (denoting senses) are substituted for names, I have not been able to derive contradictions by emulating Tarski's methods.[9] **In any case,** ***quite apart from these considerations,*** the kinds of contradictions Tarski and Montague exhibit are entirely avoided in $\mathcal{T}$-FOL by appeal to Kripke's notion of a grounded sentence. I shall expand this point in the following section, but by way of background it should prove instructive to consider briefly Perlis' approach to these difficulties.

---

[7] I have expressed the equivalence in terms of inference rules, as opposed to an axiom schema, because I believe natural deduction systems more closely approximate human reasoning.

[8] Within $\mathcal{T}$-FOL, names (constants) have senses. When a denotation is assigned to a constant, a *sense* is simultaneously assigned (logically, the sense and denotation arise together). The principle of compositionality requires that the sense of an entire sentence be composed of the senses of its parts. It follows that we cannot *assign* as the denotation of a constant the sense of a sentence which contains that constant, for the given sentence will only *have* a sense if the constant already has a sense (and thus a denotation). Clearly, I assume that it is at least possible, *temporally*, to assign senses to constants without entering a vicious cycle.

[9] Consider this example: let '*c*' name the following *sentence*. **The sense of the sentence named *c* is not true.** We may formalize this sentence as, $\neg T(sense\text{-}of(sentence\text{-}named(c)))$, where 'sense-of' and 'sentence-named' are functions. The reader may wish to amuse her/himself by attempting to derive a contradiction from this sentence by means of T1, T2, and standard FOL. Bear in mind that neither *c* nor *sense-of(sentence-named(c))* can stand alone as sentences of $\mathcal{T}$-FOL.

The key to Perlis' method is to replace Tarski's schema, True("*A*") ≡ *A*, with a modified schema:

(11) True("*A*") ≡ *A**

where *A* is a meta-variable ranging over sentences of the language, and * is an operator which ensures that (11) is equivalent to

(12) True("*A*") ≡ *A*

in those cases where '*A*' does not contain the predicate 'True'. In effect, the * operator is a device which *blocks* the derivation of True("*L*") from *L* in cases where *L* ≡ ¬*True*("*L*"), i.e., in cases where *L* is the famous "Liar" sentence. In such cases, Perlis allows that *L* may be taken as a premise, and that ¬*True*("*L*") is clearly entailed, but that *True*("*L*") may not be derived. Thus, he avoids explicit contradiction. However, although Perlis clearly demonstrates the *formal consistency* of his logic, his use of the * operator does not preclude the formation of semantically paradoxical sentences, and does not prevent the derivation of '*True(S)*' for *certain* of these paradoxical sentences, *S*. To see this, we need to have in mind Kripke's notion of a *grounded* sentence.

In essence, a grounded sentence is one whose truth evaluation "bottoms out", either immediately or ultimately, in the evaluation of sentences which do not ascribe truth or falsity to other sentences. By contrast, the truth evaluation of an *ungrounded* sentence does not "bottom out", but triggers an endless series of further attempts to evaluate the truth of itself and/or other sentences. (For more on how truth evaluation can "ultimately bottom out", see Hadley, 1989.)

Now, Perlis clearly intends that the "star" operator will block the derivation of *True("A")* from *A* when *A* is not grounded (cf., Perlis, 1985, pp. 310-311). In Perlis' schema, the "star" operator (originally due to Gilmore, 1974) works in a purely syntactic fashion; in essence, it rewrites its argument by replacing occurrences of ¬*True*("...") by occurrences of *True*("¬(...)"). But, as Kripke clearly shows, the question whether a sentence is grounded *sometimes* depends upon purely empirical considerations, and cannot reliably be determined by syntactic means. As a consequence, *certain* sentences are paradoxical in some possible worlds, but not in others. (For example, in *our* possible world, each of the following two sentences is paradoxical: *The second sentence on this page to contain a word in bold face is* **true**. *The first sentence on this page to contain a word in bold face is* **false**. Of course, in some possible worlds these two sentences would

be separated by additional sentences containing bold face, and the paradox may not arise.)

Given Kripke's conclusion, it follows that the star operator cannot do both the following things: (a) always block the derivation of *True("A")* from *A*, when *A* is ungrounded, (b) always permit (as Perlis claims) the derivation of a sentence having the form *True(...)*, from *A*, when *A* is grounded. Since the same sentence can be grounded is some possible worlds, and not in others, no purely syntactic device can reliably distinguish between semantically valid and invalid derivations of *True(...)* from ungrounded *A*.

2.3 Paradox and Truth in $\mathcal{T}$-FOL

Let us now return to the question whether paradoxes of self-reference can arise in $\mathcal{T}$-FOL. In light of the foregoing discussion of groundedness, and of the existence of *contingently* paradoxical sentences, the reader may suspect that any first-order language which permits the *ascription of truth* is vulnerable to *certain* kinds of paradox. The arguments of Kripke (1975) and Burge (1984) strongly support this proposition. Moreover, as Kripke and Burge clearly demonstrate, this vulnerability to paradox in no way presupposes syntactic or semantic ill-formedness. Of course, it is true that nearly all logics current in AI escape this vulnerability, but only because they do not sanction inferences involving truth ascriptions. For many specialized applications, one can safely ignore the logic of truth ascriptions, but if we are ever to have an adequate model of *human cognition*, we will need to represent and reason with propositions such as, "What Mary believes is true".

Since I am ultimately concerned with adequate cognitive models, I have included a truth predicate. Thus, paradoxical sentences may arise. However, Kripke (1975) has shown that the inclusion of such sentences need not render a first-order language inconsistent. Kripke proposes a complex theory of truth according to which paradoxical sentences are simply not assigned truth values. The underlying intuition of Kripke's theory is that any ascription of truth to a sentence involves a *presupposition* that the sentence is grounded. On the rare occasions when the presupposition of groundedness is unsatisfied no truth value is assigned. Kripke goes on to argue (following Kleene, 1952) that the existence of sentences without truth values does not require that we abandon any of the usual laws of logic. For example, since "$p \vee \neg p$" still holds for all cases when p *has* a truth value, we may retain the law. When p lacks a truth value, the disjunction is *not false*; rather it also

lacks a truth value. (Note that, quite apart from the existence of semantic paradoxes, there are independent reasons for believing certain well-formed sentences lack truth values. For example, Strawson (1956) has argued, and many would agree, that a sentence such as "The present king of France is bald" lacks a truth value, because the sentence involves a presupposition which is not satisfied.)

The question now arises, must the interpretation of T, the truth predicate in NIM, involve considerations of *groundedness*? In the final analysis, the answer is yes. On the face of it, the interpretation of T is straightforward. We merely assign as the extension of T the set of all senses which evaluate to 'true', where 'true' is an object in the domain of $\mathcal{L}$. However, upon reflection, we note that the relevant "set of true senses" includes not only senses of sentences taken from an agent's representation language (whatever that may be), but also senses belonging to some sentences of $\mathcal{L}$. Now, some of the latter senses arise from sentences containing the predicate T. Thus, it becomes problematic that *these* latter senses should enter into the definition of the *extension* of T (because, from an ontological standpoint, the sense and extension of a term arise together). The problem is to prevent a vicious circularity from entering into our specification of the "set of true senses" (which I denote by 'S').

My solution, in its essentials, derives from Kripke (1975).[10] Let the membership of S be *recursively* defined as follows: (In what follows, assume that ϕ is a sentence in the domain of $\mathcal{L}$. The domain of $\mathcal{L}$ includes sentences and senses belonging to the representation languages of the agents concerned, and also *all* sentences and senses belonging to $\mathcal{L}$ itself.)

(i) If ϕ contains no occurrences of T, and *if* the sense of ϕ returns 'true' on the current interpretation of ϕ, then the sense of ϕ is in S.

(ii) Otherwise, if ϕ is a sentence of $\mathcal{L}$ having the form $T(\psi)$, then the sense of ϕ (call it f) will return 'true' iff ψ is both grounded and the sense of ψ evaluates to 'true'. If f returns 'true' on the current interpretation of $\mathcal{L}$, then $f \in S$.

(iii) Otherwise, ϕ is a molecular or quantified sentence of $\mathcal{L}$, and contains some occurrence of T. The sense of ϕ (again, call it f) is formed according to the compositional rules outlined in section 2.1, and $f \in S$ iff f returns 'true' on the current interpretation of $\mathcal{L}$. Since the subexpressions of ϕ are recursively evaluated for truth in the normal Tarskian fashion, the evaluation of f ultimately reduces to conditions (i) and (ii), which are straightforward.

---

[10] In a recent work, Turner (1989) presents a solution which is in some respects similar to the solution adopted here (and which I first described in the technical report [Hadley, 1989b]).

Given the above definition, the membership of S is fully specified. This, in turn, provides a determinate extension for the predicate T. The fact that T's extension is recursively defined is no more problematic than if we had assigned the natural numbers as T's extension. As it happens, however, the chosen interpretation assures the soundness of rules of inference, T1 and T2, as we shall soon see.

2.4. Formal Properties of $\mathcal{T}$-FOL

We now consider some of the formal properties of $\mathcal{T}$-FOL. First, note that the sentences of $\mathcal{T}$-FOL, with minor syntactic alterations, may be represented in any standard first-order predicate calculus, known to be sound and complete. In what follows, 'PC' shall denote those formalizations of *pure* first-order predicate logic, which do *not* contain the propositional calculus as a subset.[11]

Definition. Let $\mathcal{T}$ be a first-order logic formed by adding to PC the two inference rules T1 and T2. Further assume that whatever Tarskian interpretation is assigned to $\mathcal{T}$, the *predicate* T is assigned the truth interpretation described in the previous section.

Proposition 1. $\mathcal{T}$ is sound.

Proof. The inference rules of $\mathcal{T}$ are exactly those of PC combined with the truth rules T1 and T2. Since the inference rules of standard PC are known to be sound, we need only demonstrate that T1 and T2 are also truth preserving. Because both these rules are explicitly concerned with truth, the proof is straightforward. In essence, T1 allows us to derive ϕ from $T(\phi)$, provided ϕ is a closed wff of $\mathcal{L}$. Now, suppose that $T(\phi)$ is true under an interpretation I, and that ϕ is a closed wff of $\mathcal{L}$. By definition of $\mathcal{T}$, I assigns the predicate T the *truth* interpretation. Furthermore, our definition of a *truth* interpretation (section 2.3), requires that $T(\phi)$ can be true only if ϕ has a true sense. But, because ϕ is a wff of $\mathcal{L}$, its sense will be true only if ϕ is true under I. Thus, T1 is truth preserving.

To see that T2 is truth preserving, we reverse the foregoing argument. T2 allows us to derive $T(\phi)$ from ϕ. Suppose that ϕ is a sentence of $\mathcal{L}$ which is true under I (as above). Then the sense of ϕ must evaluate to 'true'. Consequently, the sense of ϕ is in the extension of T. Therefore, $T(\phi)$ is true under I.

Proposition 2. $\mathcal{T}$ is complete.

The proof of completeness for $\mathcal{T}$ is complex and is given in Appendix A.

---

[11] I wish to exclude those formalizations of FOL which permit *constants* to stand alone as wffs of the language. Thus, a sentence such as $P \equiv T(P)$ does not occur in any logic denoted here by 'PC'.

2.5. Applications of the Truth Predicate

I have previously noted the importance of the concept of *truth* in human dialogue. I also claimed that rules T1 and T2 capture important aspects of common sense reasoning about truth. To support this last claim, I now consider ways in which T1 and T2 may be useful in acquiring and transmitting inductive beliefs. Suppose a robot, Bill, hears a number of Mary's remarks (call them *r1*, *r2*, *r3*, etc.) concerning the habits of waterfowl. And suppose that Bill remembers these remarks, and later has an opportunity to verify them. In time, Bill discovers that each of Mary's remarks describes what is the case. That is, Bill observes that *r1*, *r2*, *r3*.... Using rule T1, Bill infers *T(r1)* from *r1*, *T(r2)* from *r2*, etc. As time passes, Bill has a growing list of observations of the form:

$$asserts(mary, p) \wedge concerns(p, waterfowl) \wedge T(p).$$

Eventually, Bill makes an inductive leap and postulates

(13) $\forall p \{ [asserts(mary, p) \wedge concerns(p, waterfowl)] \rightarrow T(p) \}$

The use of rule T1 has been a necessary step in the formulation of Bill's generalization. Now, suppose Bill conveys (13) (in essence) to Jane, who has a growing interest in waterfowl. If Jane believes Bill, she may now regard Mary as an authority on waterfowl. When Jane is told that Mary asserted q about waterfowl, Jane may use (13) to infer *T(q)*, and then use T2 to infer q. In this manner, q may enter Jane's belief base, which in turn permits Jane to act on q, (e.g., Jane now looks for wood ducks in the first week of November). Although the preceding example by no means exhausts the ways in which we use the concept of truth, it does show that T1 and T2 have a useful role to play in everyday reasoning. Another application arises in connection with the relationship between knowledge and truth, as will emerge in the following section.

3. NIM -- A NESTED INTENSIONAL MODEL OF EPISTEMIC PROCESSES

In this section I propose NIM, an axiomatic model of epistemic processes, expressed in $\mathcal{T}$-FOL. As in $\mathcal{T}$-FOL, the present theory assumes that 'belief', 'knowledge', and 'truth' apply fundamentally to senses (which may be nested to any depth), and to sentences only derivatively. The axioms of NIM describe the processes of belief acquisition and belief change at a moderately abstract level, due to the range of topics which NIM addresses. As previously mentioned, NIM addresses the empirical

preconditions of knowledge and belief, e.g., that beliefs are acquired over time, and that conclusions drawn from existing beliefs do not automatically become new beliefs. Rather, such conclusions may first be subjected to a scrutiny process, which may or may not be successful.

NIM provides a formal conceptual map of relationships between knowledge, belief, truth, inference, questioning a belief, and retraction. Our overriding concern is not with finding a *computationally* tractable model, but with expressing the intricacies of these relationships.[12] As will become apparent, the model proposed here ('NIM' for "nested intensional model") requires an agent to know or believe *only* the senses of formulae which are explicitly present in the agent's belief base. Thus, NIM does not entail logical omniscience. However, I should caution the reader that NIM does not purport to capture *every* sense of 'belief' in common parlance. For example, there may be a usage of 'belief' according to which one believes not only *explicitly accepted* propositions, but also propositions one *would readily* accept, if asked "Do you believe this". I have here chosen not to address this *dispositional* sense of belief, because I contend that concealed in this sense are a number of contingencies which are explicitly addressed by NIM. These contingencies clearly emerge in section 4, where we consider the 'trivial-consequences' of explicitly held beliefs. (For more on the relevance of dispositional beliefs to AI, see Hadley, 1991. The work of Konolige, 1985, and Fagin & Halpern, 1985 are also relevant to this notion.)

3.1 Background Assumptions for NIM

Before presenting the axioms of NIM, it may prove helpful to make explicit a few more background assumptions. In the preceding discussion I have assumed that the objects of belief, knowledge, and truth are senses, which can be evaluated for truth. However, I shall also assume that each agent possesses a belief-base (BB) which contains a set of symbolic formulae, and that for each *believed* sense there exists some formula in the BB which explicitly (or conjuntively) expresses the believed sense in the agent's own representation language. (We say that formula f conjunctively expresses a given sense if f is a conjunction, one of whose conjuncts expresses that sense.)

---

[12]There is no reason to suppose that all relevant aspects of belief and knowledge can be expressed in tractable formal systems. We know, for example, that many forms of mathematical reasoning are intractable, and the resultant theorems are not expressible in a tractable logic. In addition, many planning problems appear insoluble, except by means of intractable search processes. If artificial agents are to engage in *hard* planning of this kind, it seems clear that they will require intractable logics. Also, if these agents are to formulate beliefs *about* how their plans were derived, it is difficult to see how these beliefs could be *represented* except in an intractable logic. For more on this, see (Hadley, 1991).

We further assume that any *premise* which the agent accepts must be represented in the BB by at least one formula. Beliefs which are *caused* in the agent via processes not normally considered inferential (e.g., via sense perception) may be regarded, for present purposes, as premises acquired at various times. In addition, the BB may contain formulae which the agent has explicitly inferred and "accepted". We place no restrictions on the particular language and inference rules which the agent X employs. If a particular agent lacks the language and axioms of NIM, we may still use NIM to reason *about* that agent's epistemic processes. Of course, an agent who possesses NIM may do the same. We do *not* assume that X's premises form a consistent set, or that X never makes a fallacious inference. However, we do assume that agents *may* employ strategies for maintaining consistency, and may, as a consequence, retract previously held beliefs. The retraction process may, or may not be ideally rational. We do not assume that an agent enters a sentence into his BB as soon as it is inferred. There is a sense of 'infer' according to which one believes anything one has inferred, but here we use 'infer' to mean merely 'derive according to accepted patterns of reasoning (which may include ''plausible reasoning'')'. Agents may treat newly inferred senses as "candidates" for belief, and enter an examination or "questioning" phase during which the candidate is tested for consistency with all or some of the agent's other beliefs.

Finally, we shall assume certain conventions throughout our discussion. For example, the expression *B(i,sense(f),t)* will frequently recur, and may be interpreted as *agent i believes the sense of formula f at time t*. (Note that '*sense*' is simply a functor in '*sense(f)*'). At times, however, we write *B(i, s, t)*. The difference lies in whether the surrounding context dictates that we are discussing formulae or senses. If we are discussing formulae, the argument to the predicates *B*, *K* and *T* must be phrased as *sense(f)*, since these predicates take senses and not formulae as arguments. In general, the axioms of NIM carefully distinguish cases where we need to refer explicitly to senses, from cases where sense arguments are implicit in the context (as in *B(i, s, t)*).

In some instances, this careful usage may seem pedantic, because in common parlance we often conflate the distinction between believing a sentence and believing what a sentence asserts. However, if NIM is to be of use to agents who conjecture about the beliefs another agent (inferentially) acquires, then it will be helpful to emphasize that beliefs involve *semantic contents* which may be expressed by distinct sentences (including sentences in different languages). For, we often speculate about what conclusions another person is drawing, even when we do not know the

precise sentences the other is using to represent his/her beliefs, thoughts, and conclusions.

3.2 Axioms of the Model

We now present the axioms of NIM. As will be apparent, NIM contains unusually many predicates (compared to the number of *operators* in a typical *epistemic logic*). This is due, I believe, to the fact that NIM deals with *aspects* of belief which are simply ignored by the logistic approach. Our contention is that one simply cannot form reasonable conjectures about an agent's belief states over time, unless these additional aspects are explicitly considered.

For clarity, the intended interpretations of each axiom and predicate are discussed as the axioms are introduced. Unless otherwise indicated, assume all variables are universally quantified with widest possible scoping.

In axiom A1 (below) the predicate *expresses(f, s)* may be interpreted as *formula f expresses sense s*, and *in-bb(i, f, t)* may be read: *f is explicitly present in i's belief base (BB) at time t.*

A1. $[expresses(f, s) \wedge in\text{–}bb(i, f, t)] \rightarrow B(i, s, t)$

Thus, A1 tells us that whenever a formula expresses a sense, and the formula is present in agent *i's* belief base, the agent believes the expressed sense. Axiom A2 (below) asserts nearly the converse, that when *i* believes *s* at *t*, there exists some formula in *i's* belief base which expresses that sense.

A2. $B(i, s, t) \rightarrow \exists f [\, expresses(f, s) \wedge in\text{–}bb(i, f, t) \,]$

The following axiom, A3, states that when two formulae, *f1* and *f2*, express an identical sense, *f1* has a sense whose belief (and knowledge) status is equivalent to that of *f2*'s sense. In essence, A3 expresses the view that synonymous formulae are interchangeable in belief and knowledge contexts.[13]

A3. $[formula(f1) \wedge formula(f2) \wedge identical(sense(f1), sense(f2))]$
$\rightarrow \{[\, B(i, sense(f1), t) \equiv B(i, sense(f2), t)] \wedge [K(i, sense(f1), t) \equiv K(i, sense(f2), t)\,]\}$

---

[13]Please recall that the criteria for synonymy can be chosen to ensure that synonymous formulae are interchangeable in belief contexts. Also, when we say that a formula is believed, we mean that *what the formula asserts* is believed. An agent may believe what a formula asserts without understanding the formula itself. Thus, on our usage, John may believe what 'Mary has a sibling asserts', even though John does not understand the word 'sibling'.

Axiom A4, below, expresses the thesis that whoever believes the sense of a conjunction must also believe the senses of its constituent conjuncts. At first glance, this thesis may seem problematic. For, one could believe (entirely on authority) that a formula (in a textbook, say) is true, without even comprehending the formula. However, *that* would not qualify as believing the sense of the formula. It is very difficult to see how one could believe the *sense* of a conjunction without believing each of its separate conjuncts.

A4. $[B(i, sense(f), t) \wedge conjunction(f) \wedge has\text{-}as\text{-}conjunct(f, conjunct)] \rightarrow B(i, sense(conjunct), t)$

The following axiom expresses the widely accepted view that whoever knows a sense s must at least believe s, and s must be true.

A5. $K(i, s, t) \rightarrow B(i, s, t) \wedge T(s)$

Note that, if we assume $K(mary, K(john, \exists y(P(y)), t1), t2)$, then we may use axiom A5, together with conjunctive simplification, to derive $T(K(John, \exists y(P(y)), t1))$. Rule T1 then allows us to obtain $K(john, \exists y(P(y)), t1)$, which in turn leads to $T(\exists y(P(y))$, via axiom A5. We may now use T1 again to derive that $\exists y(P(y))$ as a consequence of our original assumption.

Axiom A6 simply states that an agent who retracts a sense ceases to believe the sense. This, together with axiom A5, entails that the sense would not be known. Axioms A1 and A6 together entail that a retracted sense is not represented in an agent's belief base by any formula. So, although it may sound odd to speak of "retracting a sense", the oddness diminishes when we reflect that retracting a sense happens when a formula is removed from the belief base.

A6. $retracts(i, s, t) \rightarrow \neg B(i, s, t)$

In the following axiom, *questioning(i, s, t)* may be interpreted as *i begins at time t to investigate whether sense s is consistent with other beliefs*. The underlying idea is that, if i has begun to question whether s should be retained, then i is not presently committed to s's truth. From an internal standpoint, i "questions" a given sense by working with explicit formulae -- testing the consistency of a formula with other formulae, and so on. However, our concern is with the *senses* the agent questions, since we often have no access to the particular formulae the agent employs.

A7. $questioning(i, s, t) \rightarrow \neg B(i, s, t)$

We assume that those automated reasoning systems which employ *belief revision* strategies enter a phase in which the consistency of a set of senses is investigated before any retractions are made. A sense which is being questioned may be one which has been singled out for special investigation. For applications involving humans, it may be prudent to construe the *questioning* process as including some evaluation of the *interestingness* of the given proposition for the agent in question. (After all, human agents don't retain everything they infer, even when the inferred conclusion is otherwise unobjectionable.) However, I do not insist that 'retracts' and 'questioning' be interpreted just as suggested here, and no attempt is made to formalize these notions. If a given system lacks a questioning mechanism, or even a retraction mechanism, then the 'questioning' predicate (or 'retracts') may be assigned the value 'false', which would render the presence of these predicates innocuous. On the other hand, if the system has a TMS (Truth Maintainance System), the TMS will provide the means by which consistency checks may be conducted, and will provide advice about which senses are to be retracted. Thus, in AI applications, implementation of the 'questioning' and 'retracts' predicates may be comparatively straightforward.

The intended interpretation of axiom A8 (below) is that any *sense* which has just been inferred by an agent, from a subset of that agent's beliefs, is to be regarded as a candidate for belief by that agent. In the following, *just-inferred(i,s,v,t)* may be read *i inferred s from set v at time t*. *Mem(w,v)* may be read as *w is a member of set v*, and *cand(s,i,t)* is rendered as *s is a candidate for i at t*.

A8. $[just\text{–}inferred(i, s, v, t) \wedge \forall w [mem(w, v) \rightarrow B(i, w, t)]] \rightarrow cand(s, i, t)$

The following two axioms describe additional ways in which a sense may become a candidate for belief. Axiom A9 asserts that a sense which is *presented* to an agent as true (perhaps *via* sense perception) becomes a candidate for belief by the agent. Since the agent may wish to subject such a sense to a "questioning phase" before committing to belief, axiom A9 is included (cf. the remarks following axiom A12, below).

A9. $just\text{–}presented(s, i, t) \rightarrow cand(s, i, t)$

Axiom A10, below, asserts that a sense which *has* once been believed, but which is *later* being questioned, becomes merely a candidate for belief while being questioned. Axiom A7 ensures that senses are not believed while being questioned.

A10. $[B(i, s, t1) \wedge later(t1, t2) \wedge questioning(i, s, t2)] \rightarrow cand(s, i, t2)$

The following axiom tells us that, under normal circumstances, beliefs persist. More specifically, the axiom tells us that if i believes a sense at t_1, then i also believes the sense at all later times, unless i retracts or questions the sense at some intervening time. Since the interpretation of 'retracts' is left to the user's discretion, 'retracts' may be taken to include some form of spontaneous forgetting or other memory loss. Nothing here forces us to restrict retraction to deliberate actions. In A11, $later(t_1,t_3)$ may be read as t_3 is later than t_1, and $inc\text{-}between(t_1,t_2,t_3)$ may be read t_2 *lies inclusively between* t_1 *and* t_3.

A11. $[B(i, s, t_1) \wedge later(t_1, t_3)$
$\wedge \neg \exists\, t_2 \{ inc\text{–}between(t_1, t_2, t_3) \wedge [retracts(i, s, t_2) \vee questioning(i, s, t_2)] \}]$
$\rightarrow B(i, s, t_3)$

The preceding eleven axioms are intended to be *descriptive* of our concept of belief. They relate the concept of belief to cognate concepts in a manner intended to be useful to agents who reason about the beliefs of an agent. The following axiom, by contrast, is *prescriptive*. It is intended to convey the conditions under which a candidate sense (in the sense introduced by axiom A8) *should* become an active belief. The term '$successful(i, s, t_2)$' should be taken to mean that at time t_2 sense s has just been successfully scrutinized by agent i's questioning mechanism, and has been found acceptable. The term '$should\text{-}be\text{-}believed(i, s, t_2)$' may be read as, sense s should be added to i's beliefs at t_2. Thus axiom A12 states that a sense which is a candidate for i at t_1, and which has successfully passed the questioning phase at a later time t_2, should be added to i's beliefs at t_2. Of course, given A2, this would entail that a formula is added to i's belief base.

A12. $[cand(s, i, t_1) \wedge later(t_1, t_2) \wedge successful(i, s, t_2)] \rightarrow should\text{–}be\text{–}believed(i, s, t_2)$

Now, since the interpretation of 'questioning' may vary according to the intended application of the model, the interpretation of 'successful' should vary accordingly. If a particular agent altogether lacks a questioning mechanism, we may simply assign 'true' to 'successful', since there would be no tests which a candidate sense should pass, beyond being a candidate. Of course, questions may arise about how exhaustive a questioning or consistency testing process must be before we may assign 'success' to a sense. However, these are practical issues which must be decided in particular applications.

4. $\mathcal{T}$-FOL, NIM and INTER-AGENT REASONING

Throughout this discussion I have emphasized that the acquisition of beliefs by an agent presupposes that particular empirical conditions are satisfied. These conditions complicate the process by which one agent may arrive at reasonable conjectures about the belief states of *another* agent. We are rarely in a position to know *with certainty* what inferences another agent will have drawn at some future time, even when we know that agent's present beliefs. For, the inferential processes of another agent may be affected by a number of factors, including unforeseen interruptions in the other's thought processes, and "processing loads" which are presently unknown to us. Nevertheless, at least in the case of comparatively simple inferences, we are often in a position to make *plausible assumptions* about another's thought processes, and so to arrive at reasonable conclusions about particular beliefs another will acquire. In what follows we consider how $\mathcal{T}$-FOL and the axioms of NIM can help one (artificial) agent to arrive at conclusions about another agent's beliefs.

4.1 An Application of $\mathcal{T}$-FOL

To begin, we consider a problem posed by Haas (1986). Suppose that a robot (say, Robin) has the following beliefs:

(14) John knows what Mary's phone number is.

(15) John knows that Mary's phone number is the same as Bill's.

Then, at first blush, Robin is in a position to conclude

(16) John knows what Bill's phone number is.

Now, note that, strictly speaking, (16) is not a deductive consequence of (14) and (15).[14] For John

---

[14]Some may argue that (16) *is* a strict consequence of (14) and (15), on a *de re* interpretation of 'knows'. After all, if we interpret (14) as asserting that John is *acquainted* with a particular object (a number), and if (15) is true, then Mary's phone number must in fact be identical to Bill's. In that case, (16) must be true because (16) merely asserts that John is acquainted with the very same object. However, if one accepts this interpretation of 'knows', one should also accept a modification of the argument, in which (15) is replaced by "Mary's phone number is the same as Bill's". For the validity of the *de re* interpretation only requires that the identity hold, not that John knows the identity. However, I submit that most readers would find the modified argument implausible, indicating that a *de re* reading of 'knows' is unnatural in the arguments we have been considering. In what follows we assume a *de dicto* interpretation of knowledge and belief.

may have the knowledge described in (14) and (15), and be wondering what Bill's phone number is, but because John may not have it *presently in mind* that Mary's and Bill's number are the same, John may not *yet* have made the required (obvious) inference. Moreover, John's thought processes may be interrupted before John has a chance to discover the obvious. Nevertheless, it would be a reasonable assumption on Robin's part that John will make the inference, and that John will not spend time searching for inconsistencies (i.e., enter a "questioning phase"). Thus, Robin could *plausibly* adopt (16) as a working assumption. The problem arises, how do we formally reconstruct Robin's reasoning?

In his solution to this problem, Haas does not address *all* the subtleties mentioned above. Rather, he is concerned to show that Robin could represent John's knowledge in such a way that Robin could discover that the identity of Bill's phone number is entailed by what John knows. Haas argues that if Robin can discover this entailment, Robin may suppose that John would discover it as well, and make the obvious inference. So, Robin's concluding (16) would be justified (by common sense standards).

Now, given certain *idealized* assumptions, Haas succeeds in showing that Robin would be justified in concluding (16). However, as previously noted, Haas' approach to belief is purely syntactic, and involves the adoption of an elaborate quotation mechanism. As a consequence of his syntactic approach, Haas is lead to adopt a complex *reflection schema* (involving an infinity of axioms). The explanation of Haas' reflection schema is abstruse, and would take us far afield (cf. Haas, 1986, pp. 263-268). I hope to show that the intensional approach embodied in NIM leads to a comparatively simple and intuitive solution to Robin's problem. (What follows is intended as a rational reconstruction of Robin's reasoning; I do not suggest that a human reasoner would necessarily engage in the reasoning outlined below, but that reasoning similar to the following would *justify* (16) as a working assumption.)

To begin, let us consider how we might represent (14), (15) and (16) in $\mathcal{L}$. "John knows what Mary's phone number is" (i.e., (14)) becomes

(17) $\exists x\, K(john,\ eq(x,\ phone\text{–}num(mary)),\ t_a)$

where *phone-num(mary)* expresses a (grounded) function which returns the phone number of Mary,

eq is an equality predicate, and t_a is a constant denoting the time of utterance. "John knows that Mary's phone number is the same as Bill's" becomes

(18) $K(john, eq(phone\text{–}num(mary), phone\text{–}num(bill)), t_a)$

Finally, "John knows what Bill's phone number is", which is the supposed conclusion, becomes

(19) $\exists x\ K(john, eq(x, phone\text{–}num(bill)), t_a)$

We have previously noted that the supposed conclusion is not a deductively valid consequence of the given premises, and it seems clear that insofar as an inference from (14) and (15) to (16) is even *plausible*, some (trivial) inference on John's part must be presumed to take place. But John's inference requires at least *some* time, and unless an assumption is made that (14) and (15) are true *before* (16) is true, it is doubtful that we can justify Robin's concluding (16) from the two premises. It is true, of course, that in ordinary speech we often *assume* on the basis of an assertion such as (14), that John has known Mary's phone number for some *interval* of time. But even when we make assumptions such as this, the time interval implicit in (16) must begin at least slightly later than the interval implicit in (14), if Robin's inference is to be justified. We have simplified matters somewhat, by treating time intervals as isolated points of time. But, *whether or not* we adopt this simplification, the time implicit in (16) should be later than the time implicit in (14) and (15). Therefore, we should represent (16) as (20) below, and not as (19).

(20) $\exists x\ K(john, eq(x, phone\text{–}num(bill)), t_c)$

Our present assumption is that t_c (the current time of utterance) denotes a time, or time interval, later than the t_a mentioned in (17) and (18). Our goal is to demonstrate that Robin could infer (20) from (17) and (18), given some plausible assumptions that Robin may be assumed to hold concerning John. For the sake of generality, we shall pretend that the predicate *K* in (17)-(20) is replaced by *B*, since we are taking knowledge to be a special kind of belief. (Nothing crucial depends upon this; we could easily reconstruct the following arguments in terms of knowledge).

Let us assume that Robin believes (17) and (18). Robin's present task is to discover what John could (soon) come to believe on the basis of John's present beliefs. We further assume that Robin realizes that the way to explore the consequences of an agent's beliefs is to take the agent's beliefs

as premises. This entails removing embedded beliefs from their epistemic context. Because Robin is a sophisticated robot, she quickly realizes that the belief embedded in (17) can only be extracted after existential instantiation is performed. She performs the required instantiation and extracts the resulting object of belief:

(21) *eq(c, phone-num(mary))*

Because the above contains a constant introduced by existential instantiation, (21) serves as an incomplete specification of John's belief. Nevertheless, Robin can reason with (21). Likewise, after extracting the embedded belief from (18), Robin arrives at

(22) *eq(phone-num(mary), phone-num(bill))*

Assuming that Robin understands the transitivity of the '*eq*' predicate, she rapidly derives:

(23) *eq(c, phone-num(bill))*

Now, if Robin is to discover that John *could* construct a proof similar to the above, Robin must have *some* introspective abilities -- she must at least be capable of *noticing* when she herself has constructed a proof. We assume Robin has this ability, and that she can detect a trivial proof when she sees one. (For the present, we shall count those derivations as "trivial proofs" which consist of at most two immediate inferences.) On the basis of these abilities (which Robin may possess as *procedures*) she arrives at the following observation:

(24) *trivial-consequence(robin, set-of{eq(c, phone-num(mary)),*
eq(phone-num(mary), phone-num(bill))},
eq(c, phone-num(bill)),
interval$_j$)

Notice that (24) has the general form

(25) *trivial-consequence(agent, set-of(x, y), z, interval)*

Intuitively, we may render (25) as "For the agent indicated, z is a trivial consequence of the set of premises *x*, *y*, and the agent can derive *z* from {*x*, *y*} in the indicated *interval*."[15] It is important to

---

[15]We may suppose that 'set-of' expresses a function of two arguments. If both arguments are individuals which are not themselves sets, then 'set-of' returns the set of both individuals. If one argument is an individual while the other is a set, then the former is inserted into the latter, and the result is returned. If both arguments are sets, their union is returned. Using 'set-of', we can represent any set of wffs. Thus, the set {*a*, *b*, *c*} can be written as *set-of(a, set-of(b, c))*.

bear in mind that our agent, Robin, arrives at (24) through knowing the semantics of 'trivial-consequence', and through an act of introspection. Thus, (24) is not the *immediate* product of an inference any more than the observation, 'I have just derived something', is the product of inference. Also, note that the instances of *x*, *y*, and *z* in (24) are embedded wffs of $\mathcal{L}$, which return *senses* in this embedded context. Thus, (24) describes a relationship between an agent, certain senses, and a time interval. The fact that (24) identifies particular *senses*, rather than particular sentences, permits Robin to be neutral with regard to the particular sentences which Robin may suppose will occur in John's analogue to the proof of (23).

To recapitulate: we have now seen how an introspective agent (Robin) may arrive at the conclusion that a particular inference was trivial for herself. We wish to show that Robin could arrive at an analogous conclusion about another agent (John), i.e., that an inference involving the same set of senses (as are involved in Robin's inference) would be a trivial inference for John. In our own experience we find that what is truly a trivial inference for us (e.g., *modus ponens* or the transitivity of equality) is a trivial inference for anyone of even approximately equal intelligence. This is an *inductive* observation which most humans make either consciously or tacitly. We make this inductive generalization without knowing the particular reasoning mechanisms (or logics) others possess, and without knowing whether others represent facts to themselves in the *particular* way we ourselves do. If our robot is to engage in the kind of analogical reasoning humans perform, she should have the benefit of the kind of inductive knowledge humans possess. We will suppose, therefore, that Robin has among her axioms the following:

(IK1) $[equal\text{–}abilities(agent1, agent2) \wedge trivial\text{–}consequence(agent1, set\text{–}of(x, y), z, interval)] \rightarrow trivial\text{–}consequence(agent2, set\text{–}of(x, y), z, interval)$

Axiom IK1 (Inductive-Knowledge-1) simply states that when two agents have (roughly) equal deductive powers, and the first agent can derive a given consequence (by trivial means in a given interval), then the second agent can do likewise. Now, given IK1 and Robin's earlier conclusion, (24), we are in a position to show how Robin may arrive at the *plausible* conclusion that John knows what Bill's phone number is. For convenience, let us represent (24) in a shorthand fashion as:

(26) $trivial\text{-}consequence(robin, set\text{-}of(eq1, eq2), eq3, interval_j)$

In (26), *eq1* is shorthand for *eq(c, phone-num(mary))*, *eq2* for *eq(phone-num(mary), phone-num(bill))* and *eq3* for *eq(c, phone-num(bill))*. Although the plan of our proof is no doubt obvious, the proof requires some empirical assumptions on Robin's part which should be brought to light. To begin, Robin will want to use IK1. To do so, Robin needs to believe that John's deductive abilities are about equal to her own. Let us assume that Robin has such a belief, perhaps acquired through observation. So, in addition to (26) (and other beliefs) Robin believes

(27) *equal-abilities(robin, john)*

Using (26), (27) and IK1 Robin infers

(28) *trivial-consequence(john, set-of(eq1, eq2), eq3, interval$_j$)*

Now, given that Robin has deduced that *eq3* is a trivial consequence (for John) of John's beliefs, we might suppose that Robin will continue to reason as follows: Since *eq3* is so easily derived from *eq1* and *eq2*, John will in fact derive *eq3* from existing beliefs. This would make *eq3* a *candidate* in the sense introduced by NIM axiom A8. Once *eq3* becomes a candidate, it is unlikely that John would subject it to a *questioning* phase, because *eq3* is such a trivial consequence of existing beliefs. So, *eq3* would become a *successful* candidate, by default. At this point, Robin could use NIM A12 to derive that *eq3* should have become one of John's beliefs by a certain time.

However, although Robin *might* reason as above, this account of Robin's reasoning assumes that Robin is in possession of (something equivalent to) NIM's axioms. While I think it likely that humans hold *tacit* beliefs equivalent to NIM's axioms, we need not suppose that agents actually *use* most of these beliefs when reasoning about another's beliefs. By analogy, a physicist does not typically use the laws of physics to reason about everyday movements in the world. Rather, she/he relies upon *inductively* acquired, qualitative knowledge when moving everyday objects in the world. When such approximate, qualitative knowledge fails, physical laws may be invoked to explain the failure. Similarly, we will suppose that agents use naive inductive "knowledge" to reason about beliefs other agents inferentially acquire. When this "knowledge" fails, NIM axioms may be used to explain the failure.

The question arises, what inductive knowledge will allow Robin to move from (28) to the desired conclusion? Well, presumably, Robin has had sufficient experience with trivial inferences in her

own case to have observed that conclusions resulting from these inferences are typically believed. Moreover, Robin has made similar observations in the case of other agents -- people very often adopt the *immediate* consequences of their beliefs as new beliefs. It seems plausible, therefore, that Robin should have an inductive belief equivalent to:

(IK2) {*trivial–consequence(agent, set–of(x, y), z, interval)*
$\wedge$ $\forall$ *w*[*mem(w, set–of(x, y))* $\rightarrow$ *B(agent, w, t1)*]
$\wedge$ *follows(t1, interval, t2)* }
$\rightarrow$ *moderately–likely(B(agent, z, t2))*

Principle IK2 tells us that, for a given *agent* and *interval*, if *z* is a trivial consequence of *set-of(x, y)*, and the agent believes every member of *set-of(x, y)* at *t1*, and time *t2* follows *t1* by the given interval, then it is moderately likely that the agent believes *z* at *t2*.

Now, given Robin's prior conclusion (28) and that Robin realizes that John believes both *eq(c, phone-num(mary))* and *eq(phone-num(mary), phone-num(bill))* at t_a (i.e., John believes every member of *set-of(eq1, eq2)* at time t_a), and given further that Robin knows that t_c follows t_a by $interval_j$, it is clear that Robin can derive

(29) *moderately-likely(B(john, eq(c, phone-num(bill), t_c)))*

by an application of universal instantiation and *modus ponens* to IK2. Assuming that Robin arrives at (29) in this fashion, Robin is entitled (at least by common sense standards) to adopt (30) as a working assumption:

(30) *B(john, eq(c, phone-num(bill), t_c))* .

From (30) Robin may easily obtain the desired (working) conclusion:

(31) $\exists$ *x B(john, eq(x, phone-num(bill), t_c))*.

It is worth stressing that in adopting (30) as a working assumption, Robin is not categorically committed to the truth of (30), and *should* retain a mental note of (30)'s uncertainty. Adopting (30) as a working assumption merely allows one to keep a *separate* record of confidence ratings (such as 'moderately-likely'), while permitting one to work with the propositions so rated.

Perhaps the objection will arise that it ought to be possible to derive something stronger than (29)

and, consequently, something stronger than the working assumptions (30) and (31). Against this objection I would argue that an overriding concern of the present work has been to acknowledge the uncertainties introduced by the following considerations: (a) agents frequently fail to notice even the immediate consequences of their beliefs, (b) agents sometimes abandon such consequences by modifying their premise set. It is only by ignoring such complications that Haas is able to derive something stronger than (29) from (14) and (15). I contend that concerns for cognitive validity and resource limitations drive us towards conclusions such as (29) and away from idealized conclusions, such as (16).

In passing, we should note that the ability to adopt working assumptions such as (30) can be important in AI systems which interact with human users (e.g., automated tutoring or question-answering systems). For, it can be tedious for human users of such systems to constantly affirm the new beliefs they have acquired via trivial inferences. Equally important, however, is that the AI system record the uncertainty associated with such working assumptions in order that uncertainties can be propagated and compounded, when *new* beliefs are attributed on the basis of prior working assumptions (such as (30)). Also, note that as long as uncertainty ratings are compounded through propagation, we are in no danger of attributing *logical omniscience* to other agents. For, compounded uncertainties rapidly diminish to the point where it would be reckless to base working assumptions upon the resultant certainty values. (cf. Aleliunis, 1988, for a discussion of qualitative probability theory, and Pearl, 1986, for a discussion of propagation of probabilities within belief networks.)

4.2 A Diagnostic Application of NIM

We have now seen how the language of $\mathcal{T}$-FOL, together with two inductive principles (IK1 and IK2), enables a limited form of inter-agent reasoning. I have also briefly outlined how NIM can be used to justify such reasoning. Let us now consider how NIM can provide an automated system (say, Robin) with a (formal) conceptual framework for uncovering mistaken assumptions in the kind of reasoning exhibited above. Suppose that a human, Pam, is "conversing" with Robin about Pam's finances. (We assume that Robin contains, among other things, the natural language component of a *financial expert system*.) Initially, Pam supplies Robin with a base set of Pam's relevant beliefs, so that Robin may form an initial "model" of Pam. Now, suppose that after a brief interchange, Robin projects onto Pam a new belief, which Robin supposes to have resulted

from a trivial inference on Pam's part. Thus, Robin has adopted a working assumption of the form:

(32) *B(pam, s1, t2).*

But, suppose further that after additional "conversation" with Pam, Robin has good reason to doubt (32). Given that Robin only regards (32) as *moderately-likely*, she is in a position to abandon (32), but may still seek to explain just how (32) could be false, given her assumptions about Pam's other beliefs. If Robin has been endowed with the NIM axiom set, two possible explanations should emerge. To see this, consider the NIM axioms A8 and A12.

A8. $[\mathit{just\text{–}inferred}(i, s, v, t) \wedge \forall w \{ \mathit{mem}(w, v) \rightarrow B(i, w, t) \}] \rightarrow \mathit{cand}(s, i, t)$

A12. $[\mathit{cand}(s, i, t_1) \wedge \mathit{later}(t_1, t_2) \wedge \mathit{successful}(i, s, t_2)] \rightarrow \mathit{should\text{–}be\text{–}believed}(i, s, t_2)$

We assume that Robin "accepts" the NIM axioms, and thus understands that every change in an agent's *belief set*, over time, is explicable in terms of a derivation within NIM (together with background conditions). So, if Robin's initial conjecture that (32) were correct, it must be possible to derive (32) within NIM, given Pam's initial beliefs, and information about what inferences, questionings, retractions, etc., Pam has made. Robin's strategy, therefore, ought to be to discover what assumptions, over and above Pam's initial beliefs, would be required to derive (32) within NIM.

Given any reasonable proof planning abilities, and *some operational* understanding of the predicate, *should-be-believed*,[16] Robin will be led to axiom A12, which tells us that *s1* would need to be both a *candidate* and *successful* before becoming a belief. Now, again, we must assume Robin has been given *some operational* understanding of the predicate *successful* (in terms of computational, consistency-checking mechanisms). Given this, Robin should rapidly discover that she has no real evidence that Pam ever found *s1* to be *successful*. Moreover, Robin will soon discover that *s1* has never been established as a candidate. Robin's concern for *trivial-consequence* should lead her to A8, which *would* establish *s1* as a candidate if it could be shown that at some

---

[16]Although it is well known that most natural language predicates cannot be given operational definitions, NIM axioms A1 and A2 provide Robin with a *formal* basis for assuming that believed senses are represented explicitly in an agent's belief base. A predicate such as '*should-be-believed*' could tell Robin to assume the given sense will be so represented. We also assume that predicates such as 'questioning' and 'successful' can be operationally defined *to some degree*, for example, by being associated with computational processes (as evidenced by the existence of TMSs).

point *s1* had just been inferred. At this point, Robin *should* discover that, although she has established *s1* as a trivial-consequence of Pam's prior beliefs, she has no evidence the Pam actually inferred *s1*. Thus, given *some* operational understanding of the relevant predicates, together with standard goal-reduction techniques, Robin should arrive at the following possibilities: either (32) is false because Pam did not infer *s1* before *t2*, or Pam did infer *s1*, but *s1* was not "successful".

We have just seen how an AI system endowed with NIM may discover ways in which normal inductive reasoning about another's belief states can go astray. Such an ability is important, because it may assist the system in uncovering useful regularities in the behaviour of other agents. For example, humans are more likely to overlook obvious implications of their beliefs (or to *resist* these implications by rejecting a hitherto believed premise) in some domains than in others. The domain of personal finances may be a case in point. Our system, Robin, may discover that people frequently *reject* the consequences of their beliefs when those consequences indicate the need for greater financial austerity. In this way Robin may fine-tune her inductive principles.

In addition to the applications so far considered, the axioms of NIM provide an AI system with a formal mapping of important interrelationships between belief-bases, senses, formulae, retraction, knowledge, truth, questioning, and the like. Moreover, these interrelationships are temporally defined. Although we do not suppose any existing AI system to be capable of *truly understanding* these relationships, possession of a formal mapping should assist the system in playing the appropriate "language game" with the system user. For example, axiom A3 informs the system that distinct sentences can express the same belief. The system may then infer that when an agent has a belief, the agent's belief-base need not contain a representation identical to the system's own representation. Axioms A1 and A2 enable the system to discover the relationship between having a belief and having stored some representation of the belief. Axiom A5 informs the system of certain *necessary* conditions for knowing something,[17] while axioms A6-A12 display interrelationships between 'believing', 'questioning', 'candidate', and 'retraction'.

5. SUMMARY

[17]The problem of formulating *sufficient* conditions for 'knowledge' has occupied philosophers since the time of Socrates. Although considerable progress has been made in the last thirty years, no consensus on sufficient conditions yet exists.

In the foregoing I have presented a representation language, logic, and a theory of epistemic processes which cooperatively solve a family of related problems. Section 2 describes the language and logic, $\mathcal{T}$-FOL, which simultaneously permits the representation of truth ascriptions, implicit self-reference, and arbitrarily embedded sentences within a first-order setting. Through the combined use of principles derived from Frege, Montague, and Kripke, together with context-sensitive semantic conventions, $\mathcal{T}$-FOL captures the logic of truth inferences, while avoiding the inconsistencies exhibited by Tarski. The soundness and completeness of $\mathcal{T}$-FOL have been established herein. By virtue of a distinguished truth predicate, and the inclusion of inference rules T1 and T2, $\mathcal{T}$-FOL attains an expressive power approaching that of natural language. We have seen that knowledge of how to reason about *truth* permits agents to acquire and transmit useful inductive knowledge.

In section 3, a sense-based, *process* model of belief and knowledge was presented, within the language of $\mathcal{T}$-FOL. The model, NIM, sanctions inferences involving nested epistemic attitudes, with different agents, and different times. Moreover, NIM's *process-oriented* focus makes explicit many (if not all) of the empirical preconditions of belief acquisition. Thus, NIM underscores the limitations of inter-agent reasoning. Equally important, NIM provides agents with a conceptual map, interrelating the concepts of knowledge, belief, truth, and a number of cognate concepts, such as 'infers', 'retracts', and 'questions'.

Section 4 describes applications of $\mathcal{T}$-FOL and NIM to inter-agent reasoning. $\mathcal{T}$-FOL's capacity for representing nested propositional attitudes, and for "quantifying into" epistemic contexts, permits one agent to reason conjecturally about another's epistemic processes, without knowing the other's representation scheme. This conjectural reasoning is facilitated by two factors: (a) $\mathcal{T}$-FOL's sense-based semantics, and (b) the fact that an agent can arbitrarily embed propositions within predicates such as 'moderately-likely', 'doubtful', etc. In addition, when such conjectural reasoning goes astray, NIM supplies an automated system with an axiomatic framework in which to diagnose the sources of error, much as a detailed physical theory can be used to diagnose erroneous reasoning which arises from inductively acquired, qualitative principles. Such diagnostic abilities may be important to automated systems which engage human users in sophisticated dialogue.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the support of the Natural Sciences and Engineering Council of Canada (Grant No. A0899) during the period of this research.

My sincere thanks to the following for discussions which stimulated many of the thoughts contained herein: James Delgrande, Randy Goebel, Raymond Jennings, Charles Morgan, Brian Cantwell Smith, Lenhart Schubert and Susan Wendell. Also, I am indebted to Raymond Jennings, Charles Morgan, and Susan Wendell for their helpful comments on an earlier version of this paper.

Appendix A

Proof of the completeness of $\mathcal{T}$.

The language of $\mathcal{T}$ is simply the language of PC. No new vocabulary has been introduced, although the interpretation of T is fixed. Moreover, the interpretation of T is entirely consistent with the principles of Tarskian interpretation. Now, the inference rules of FOL are a subset of $\mathcal{T}$, and these rules are known to be complete for any first-order theory. Thus, these rules ensure that every *universally valid* consequence of $\mathcal{T}$ is derivable in $\mathcal{T}$, since $\mathcal{T}$ is just a first-order theory with additional rules of inference. The two additional rules do not block the application of any of the standard FOL inference rules. However, since we are restricting the models of $\mathcal{T}$ to those where T has a special interpretation, we must consider whether any *additional* consequences hold in these *restricted* models (hereafter, T-models) which do not hold in the unrestricted set of all models for $\mathcal{T}$. Clearly, the answer is yes. In particular, we know that any T-model which satisfies a sentence of the form T(ϕ) will also satisfy ϕ, and any T-model which satisfies ϕ will satisfy T(ϕ).

Now, to establish the completeness of $\mathcal{T}$ we need to show that if Γ is a set of sentences expressed in $\mathcal{T}$, and ϕ is a consequence of Γ, then ϕ is derivable from Γ. "Consequence of Γ" here includes all sentences whose truth is *semantically entailed* by the joint assumption of Γ with the special interpretation of T as a truth predicate. In general outline, our proof follows Mate's [1965] version of the standard Henkin completeness proof for FOL. However, modifications are required to the standard proof in order to accommodate our fixed interpretation of T. These modifications will be explained in detail. In the following it is important to bear in mind that the language of $\mathcal{T}$, which is $\mathcal{L}$, contains only the syntax and vocabulary of traditional FOL with functions. The proof of

completeness requires the following major lemma, from which completeness follows by a standard, straightforward argument.

Lemma I. If a set of sentences Γ (of $\mathcal{L}$) is d-consistent, then Γ is consistent.

Customarily, Lemma I is proved by means of a special case of I, namely:

Lemma I′ If Γ is a d-consistent set of sentences, and every individual constant occurring in a sentence of Γ has an *even* index (e.g., *c2*, *c4*, *etc.*), then Γ is consistent.

Once I′ is established, Lemma I follows immediately, since those Γ which satisfy the antecedent of I′ are merely notational variants of those which satisfy the antecedent of I. Now, I′ can be established as an immediate consequence of the following two lemmas:

Lemma II. If Γ is a set of sentences satisfying the antecedent of I′, then there is a maximal d-consistent, ω-complete set of sentences, Δ, such that $\Gamma \subseteq \Delta$.

Lemma III. If Δ is maximal d-consistent and ω-complete, Δ is consistent.

The proof of II is constructive, and requires the existence of an infinite set of individual constants not mentioned in Γ. However, the antecedent of I′ was designed to ensure that all constants having *odd* indices do not occur in Γ and are available for use in the construction. The traditional proof of II involves the construction of a maximal d-consistent, ω-complete set Δ which contains Γ. This construction requires that the sentences of $\mathcal{L}$ can be enumerated in a particular sequence, and that the usual rules of FOL be available. However, nothing in the theory $\mathcal{T}$ affects this construction, since the language of $\mathcal{T}$ is identical to FOL and neither of the rules T1 or T2 interferes with the application of the standard rules of FOL. The reader may readily verify that the standard proof of II is unaffected by our fixed interpretation of the predicate *T*. For this reason, the proof of II is omitted.

Proof of III.

Let Δ be an ω-complete, maximal d-consistent set of sentences of $\mathcal{L}$. We specify an interpretation *I*, for $\mathcal{L}$, as follows. The domain of $\mathcal{L}$ will be the set of all sentences, and senses of sentences of $\mathcal{L}$. Any given individual constant, α, of $\mathcal{L}$, will denote the sense of a unique tautology, $pred(\alpha) \equiv pred(\alpha)$, *provided* Δ contains the sentence $T(\alpha)$; otherwise, α denotes the sense of the necessarily false sentence, $pred(\alpha) \equiv \neg pred(\alpha)$. Each grounded *functional* expression, β, which (when removed from its "embedded" context) would *not* constitute a well-formed sentence of $\mathcal{L}$, will denote either the sense of a unique tautology involving β, or the sense of a necessarily false sentence involving β, depending upon whether $T(\beta)$ is in Δ (as in the case of α, above). Each

embedded functional expression which *could* "stand alone" as a sentence of $\mathcal{L}$ has as denotation its *own sense*. To each n-ary predicate π of $\mathcal{L}$ (not identical to the unary predicate T), I assigns just those n-tuples of objects in the domain of $\mathcal{L}$, $<\delta_1, \ldots . \delta_n>$, such that there exists an *atomic* sentence in Δ having the form $\pi(\gamma_1 \ldots . \gamma_n)$, where each term in the n-tuple of expressions ($\gamma_1 \ldots . \gamma_n$) denotes a corresponding object in the n-tuple $<\delta_1, \ldots . \delta_n>$. To the *unary* predicate T, I assigns (recursively, as in section 2.3) the set of senses of those sentences of $\mathcal{L}$ which are true under I.

We now define the *index* of expressions of $\mathcal{L}$ as follows: The index of any sentence or *functional expression* of $\mathcal{L}$ shall be the total number of quantifiers, connectives, and Ts occurring in that sentence, including those occurrences embedded in its (functional) arguments. Thus, the atomic sentence $T(\exists x (T(x)))$ has an index of three, $\exists x (pred(x))$ has index one and $pred(c)$ has index zero.

Now, let the sentences of $\mathcal{L}$ be divided into two disjoint sets. Set S_N contains all atomic sentences *except* those having the form $T(\psi)$, where ψ contains one or more predicates. Thus, $T(b)$ is in S_N, but $T(pred(b))$ is not. Set S_T contains *all other* sentences of $\mathcal{L}$.

We must now prove that

(C)

Every sentence ϕ of $\mathcal{L}$ is true under I (I-true) iff $\phi \in \Delta$.

(Note, the expression 'sentence ϕ is true under I' should be taken as elliptical for 'the sense of sentence ϕ is true under I'.) We begin by proving a special case of (C), namely

(C')

Every sentence ϕ of S_N is true under I iff $\phi \in \Delta$.

Now, clearly, by the construction of I, all true (atomic) sentences in S_N are in Δ, and those atomic sentences in Δ which are also in S_N are true. This is because I was constructed so that every predicate π in $\mathcal{L}$, not equal to unary T, has in its denotation just those n-tuples of objects which are denoted by the n-tuples of expressions which occur as arguments to the π predicates in Δ. Moreover, those atomic sentences in S_N having the form $T(\alpha)$ satisfy (C'), because the

denotations of all (non-predicate containing) arguments to *T* were chosen so as to ensure this.

We have now established (C'). To establish (C) it remains to show that all sentences of S_T satisfy (C). Now, since every sentence in S_T is either non-atomic or contains the predicate *T*, it follows that the index of every sentence in S_T is > 0. Let us suppose that (C) fails to hold for some sentence in S_T. Let ϕ be such a sentence in S_T having lowest index, N. Then ϕ must have one of the eight forms listed below.

Case (1) $\phi = T(\psi)$, where the index of $\psi < N$.

Now, by the definition of the extension of *T* we have that

ϕ is *I*-true iff ψ is *I*-true (because ψ is *I*-true iff the sense of ψ returns 'true' under *I*, which happens just when the sense of ψ is in the extension of *T*.)

iff $\psi \in \Delta$ (since the index of $\psi < N$, and the induction hypothesis is assumed.)

But, ψ can be in Δ iff ϕ is also, because Δ is d-consistent and ψ and ϕ are each derivable from the other, by T1 and T2. It follows that *if* ϕ is a sentence of $\mathcal{L}$ which fails to satisfy (C), then ϕ cannot have the form $T(\psi)$, and case (1) is therefore excluded. Note that we have now established that *every* atomic sentence of $\mathcal{L}$ satisfies (C), for we have proven (C) both in the earlier case where $\phi \in S_N$, and in the present case where $\phi \in S_T$ and ϕ is atomic. No other atomic sentences remain.

Cases (2-6) have the forms listed below:

(2) $\phi = \neg(\psi)$
(3) $\phi = \psi \vee \chi$
(4) $\phi = \psi \wedge \chi$
(5) $\phi = \psi \rightarrow \chi$
(6) $\phi = \psi \equiv \chi$

The proofs for the above five cases follow the usual Henkin method. Since these cases do not differ from those found in standard completeness proofs for FOL, the proofs for these cases are omitted.

(7) $\phi = \forall \alpha (\psi)$
(8) $\phi = \exists \alpha (\psi)$

In cases (7) and (8) we must depart from the standard induction process for the following reason. Although the index of ψ is < N, not all ground instances of ψ have index < N. This is because

occurrences of α in ψ may be replaced by a functional expression having an index > 0. For this reason we now adopt a slightly different strategy. We shall prove, independently of the (previously defined) indexing scheme, that all instances of (7) and (8) satisfy (C), and therefore that the supposed counter-example could not belong to either of these cases.

To begin, recall that we have already established that every atomic sentence of $\mathcal{L}$ satisfies (C). This fact enables us to introduce a notion of *standard* index, defined as follows: all atomic sentences have a standard index of zero (regardless of the complexity of arguments appearing in the atomic sentence). The standard index of any other sentence is simply the total number of quantifiers and logical connectives which occur in that sentence at levels *above the atomic level.* Using this concept of standard index we establish that *regardless* of the non-standard (previously defined) index of ϕ, ϕ must satisfy (C) if ϕ belongs to case (7) or (8). For, suppose that ϕ does not satisfy (C). Then there must exist some sentence χ (having form (7) or (8)) which fails to satisfy (C), and which has the lowest *standard* index of any such sentence. Call this index M. Now, if χ belongs to case (7), then we have $\chi = \forall\, \alpha(\Sigma)$, where the standard index of Σ is M-1.

Note that, since all atomic sentences have a standard index of zero, all instantiations of Σ (i.e., instances obtained by instantiating on just the α occurrences in Σ) also have index less than M. (Bear in mind that replacing α in Σ by a complex expression could not increase the standard index of any atomic wff within Σ which contained α.) Hence, by hypothesis, all instances of Σ satisfy (C), i.e., each such instance is in Δ iff that instance is *I*-true. Now suppose that χ is true. Then all of Σ's instances are true, and so are in Δ. But, since Δ is ω-complete, χ must be in Δ.

On the other hand, suppose χ is in Δ. By ω-completeness, all instances of Σ are in Δ. Since their index is less than M, these instances must all be true. Hence, χ is also true (by definition of $\forall$). Therefore, we have that $\chi \in \Delta$ iff χ is true, contrary to the assumption that χ violates (C). It follows that every sentence of $\mathcal{L}$ having the form of (7) satisfies (C). The argument for case (8) is precisely analogous to that for (7). We conclude, therefore, that the supposed sentence of $\mathcal{L}$ which fails to satisfy (C) cannot fall into case (7) or (8). Since no other cases remain, the supposed counter-example does not exist, and (C) holds for all sentences of $\mathcal{L}$. Moreover, since (C) assumed only a specific interpretation for the expressions of $\mathcal{L}$, and that Δ is d-consistent and ω-complete, we have shown that every such Δ is satisfiable. This completes the proof of III.

REFERENCES

Aleliunis, R., (1988) "A summary of a new normative theory of probabilistic logic", Proceedings of the 4th Workshop on Uncertainty in AI, University of Minnesota, pp. 8-14.

Barnden, J.A., (1986) "Imputations and explications: representational problems in treatments of propositional attitudes", *Cognitive Science*, Vol. 10, pp. 319-364.

Barwise, J. & Perry, J., (1983) *Situations and Attitudes*, Cambridge, MA: Bradford Books.

Brachman, R. J., (1977) "What's in a concept: structural foundations for semantic networks," *International Journal of Man-Machine Studies*, Vol. 9, pp. 127-152.

Burge, T., (1984) "Epistemic paradox", *Journal of Philosophy*, Vol. 81, pp. 5-29.

Church, A. (1954), "Intensional isomorphism and identity of belief", *Philosophical Studies*, 5, pp. 65-71.

Cresswell, M.J., (1985) *Structured meanings: the semantics of propositional attitudes*, Cambridge, MA: MIT Press.

de Kleer, J., (1986) "An assumption-based TMS", *Artificial Intelligence*, Vol. 28, pp. 127-162.

Doyle, J., (1979) "A truth maintenance system", *Artificial Intelligence*, Vol. 12, pp. 231-272.

Doyle, J., (1983) "The Ins and Outs of Reason Maintenance", in Proceedings of IJCAI (International Joint Conference on Artificial Intelligence), Karlsruhe, Germany, pp. 349-351.

Fagin, R., and Halpern, J., (1985) "Belief, awareness, and limited reasoning", Proceedings of IJCAI, Los Angelos, pp. 491-501.

Frege, G., (1892), "On Sense and reference", in *Translations from the Philosophical Writings of Gottlob Frege*, (Eds.) Geach, P. & Black, M., (1952) Oxford: Basil Blackwell.

Gilmore, P., (1974) "The consistency of partial set theory without extensionality", in *Axiomatic Set Theory*, (Ed.) Jech, T., Providence, RI: American Mathematical Society.

Goodwin, J.W., (1985) "A process theory of non-monotonic inference", Proceedings of the Ninth International Joint Conference on Artificial Intelligence, August, Los Angeles, pp. 185-187.

Haas, A., (1985) "Possible events, actual events, and robots". *Computational Intelligence*, Vol. 1, No. 2, pp. 59-70.

Haas, A., (1986) "A syntactic theory of belief and action", *Artificial Intelligence*, Vol. 28, 245-292.

Hadley, R.F., (1986) "Fagin and Halpern on logical omniscience, a critique with an alternative", Proceedings of the Sixth Canadian Conference on Artificial Intelligence, Montreal, May, pp. 49-56.

Hadley, R.F., (1988) "Logical omniscience, semantics and models of belief", *Computational Intelligence*, Vol. 4, pp. 17-30.

Hadley, R.F., (1989) "A default-oriented theory of procedural semantics", *Cognitive Science*, Vol. 13, pp. 107-137.

Hadley, R.F., (1989b) "An intension-based model of nested epistemic processes", Technical Report CSS-IS TR 89-04, Centre for Systems Science, Simon Fraser University, Burnaby, B.C., Canada.

Hadley, R.F., (1991) "The many uses of 'belief' in AI", *Minds and Machines*, Vol. 1, (in press).

Halpern, J. & Moses, Y., (1985) "A guide to the modal logics of knowledge and belief: preliminary report", Proceedings of IJCAI, Los Angelos, pp. 480-490.

Hobbs, J.R. & Rosenschein, S.J., (1978) "Making Computational Sense of Montague's Intensional

Logic", *Artificial Intelligence*, 9, pp. 287-306

Kleene, S.C., (1952) *Introduction to meta-mathematics*, New York: Wolters-Noordhoff Publishing and North-Holland Publishing.

Konolige, K., (1985) "A computational theory of belief introspection", Proceedings of IJCAI, Los Angelos, pp. 503-508.

Kripke, S., (1975) "Outline of a theory of truth", *Journal of Philosophy*, Vol. 72, pp. 690-716.

Lakemeyer, G., (1987) "Tractable meta-reasoning in propositional logics of belief", Proceedings of IJCAI, Milan, pp. 401-408.

Levesque, H.J., (1984) "A logic of implicit and explicit belief", Proceedings of the National Conference on Artificial Intelligence, Austin, Texas, pp. 198-202.

Lewis, D., (1970) "General semantics", *Synthese*, Vol. 22, pp. 18-67.

Maida, A. S., & Shapiro, S. C., (1982) "Intensional Concepts in Propositional Semantic Networks," *Cognitive Science* 6, pp. 291-330

Maida, A.S., (1986) "Introspection and reasoning about the beliefs of other agents", *Proceedings of the Cognitive Science Society*, pp. 187-195.

Martins, J., & Shapiro, S., (1988) "A Model of Belief Revision", *Artificial Intelligence*, Vol. 35, pp. 25-79.

Montague, R., (1960) "A paradox regained", reprinted in *Formal Philosophy*, (Ed.) Thomason, R., New Haven, CN: Yale University Press, (1974), pp. 271-285.

Montague, R., (1970) "Universal grammar", *Theoria* 36, pp. 222-246.

Moore, G.E., (1942) "Reply to my critics", In *The philosophy of G.E. Moore*, (Ed.) P. Schilpp., Evanston and Chicago: Northwestern University Press, pp. 660-667.

Pearl, J., (1986) "Fusion, propagation, and structuring in belief networks", *Artificial Intelligence*, Vol. 29, pp. 241-288.

Perlis, D., (1985) "Languages with self-reference I: foundations", *Artificial Intelligence*, Vol. 25, pp. 301-322.

Perlis, D., (1988) "Languages with self-reference II: Knowledge, Belief, and Modality", *Artificial Intelligence*, Vol. 35, pp. 179-212.

Rapaport, W. J., (1986) "Logical Foundations for Belief Representation," *Cognitive Science*, 10, pp. 371-422.

Strawson, P.F., (1956) "On referring", in *Essays in Conceptual Analysis*, (Ed.) Flew, A., London: MacMillan and Co. Ltd.

Tarski, A., (1936) "Der warheitsbegriff in den formalisierten sprachen", *Studia Philos.*, Vol. 1, pp. 261-405.

Turner, R., (1989) *Truth and Modality for Knowledge Representation*, Pitman Press, London.

Wilks, Y. & Ballim, A., (1987) "Multiple agents and the heuristic ascriptions of belief", Proceedings of IJCAI, Milan, pp. 118-124.

Centre for Systems Science

• Simon Fraser University established the **Centre for Systems Science** as a provincial centre for advanced study. The objective of CSS is twofold: coordinate and combine advanced systems research at SFU, and encourage application of research results within industry in order to further the technological development of British Columbia.

• CSS has assembled three research groups to streamline its efforts. The orientation of these groups is as follows: intelligent systems, microelectronics, and computer and communications systems. Because of shared research interests, this latter group merges with LCCR (**Laboratory for Computer and Communications Research**) for the publication of technical reports.

• CSS's three groups involve over 80 skilled faculty researchers principally from SFU's Schools of Engineering Science and Computing Science, but also draw substantially on Mathematics, Physics, Geography, Philosophy, Linguistics, Psychology, Communication, and Kinesiology.